JOURNAL OF CONSULTING PSYCHOLOGY

Edited By
Laurance F. Shaffer
Teachers College, Columbia University

Advisory Editors

Donald W. Fiske
University of Chicago

Ann Magaret Garner
University of Illinois

Harrison G. Gough
University of California

William A. Hunt
Northwestern University

E. Lowell Kelly
University of Michigan

Morris Krugman
Board of Education
New York, N. Y.

Bertha M. Luckey
Board of Education
Cleveland, Ohio

Fred McKinney
University of Missouri

Anne Roe
New York, N. Y.

Carl R. Rogers
University of Chicago

R. Nevitt Sanford
University of California

Edward Joseph Shoben, Jr.
Teachers College
Columbia University

George J. Wischner
University of Pittsburgh

J. R. Wittenborn
Rutgers University

Managing Editor
Lorraine Bouthilet

VOLUME 20
1956

BUREAU OF EDNL. & PSYL. RESEARCH
LIBRARY

Published Bimonthly By

The American Psychological Association, Inc.
1333 Sixteenth St., N. W.
Washington 6, D. C.

Printed At
Lancaster, Pennsylvania

Copyright © 1956 by the American Psychological Association, Inc.

Contents of Volume 20

Altus, Grace T. A WISC Profile for Retarded Readers - - - 155

Altus, Grace T. Some Correlates of the Davis-Eells Tests - - - 227

Ansbacher, H. L. The "Alienation Syndrome" and Adler's Concept of "Distance" - - - 483

Barnes, E. H. Factors, Response Bias, and the MMPI - - - 419

Barnes, E. H. Response Bias and the MMPI - - - 371

Barratt, E. S. The Relationship of the Progressive Matrices (1938) and the Columbia Mental Maturity Scale to the WISC - - - 294

Beck, L. H. See *Rosvold, H. E.*

Bellugi, Ursula. See *Crandall, V. J.*

Bendig, A. W. Ranking Methodology: the Development of a Judgmental Criterion with Clinical Case Histories - - - 75

Bendig, A. W. Reliability of Case History Ratings and Intellectual Ability of Graduate Raters - - - 142

Bendig, A. W. The Development of a Short Form of the Manifest Anxiety Scale - - - 384

Bendig, A. W. The Personality of Judges and their Agreement with Experts in Judging Clinical Case Histories - - - 422

Berg, J., and Polyot, C. J. The Influence of Color on Reactions to Incomplete Figures - - 9

Bernstein, L. The Examiner as an Inhibiting Factor in Clinical Testing - - - 287

Bieri, J., and Blacker, E. External and Internal Stimulus Factors in Rorschach Performance - - - 1

Blacker, E. See *Bieri, J.*

Blum, G. S. "Reliability of the Blacky Test": A Reply to Charen - - - 406

Bordin, E. S. See *Raush, H. L.*

Bosquet, K. T., and Stanley, W. C. Discriminative Powers of Rorschach Determinants in Children Referred to a Child Guidance Clinic - - - 17

Bransome, E. D., Jr. See *Rosvold, H. E.*

Breiger, B. The Use of the W-B Picture Arrangement Subtest as a Projective Technique - 132

Breiger, B. See *Lundin, W. H.*

Brody, Janice R. See *Levy, L. H.*

Broen, W. E., Jr. See *Wirt, R. D.*

Calvin, A. D., McGuigan, F. J., Tyrrell, Sybil, and Soyars, M. Manifest Anxiety and the Palmar Perspiration Index - - - 356

Carpenter, B. See *Wiener, M.*

Carpenter, Janeth T. See *Wiener, M.*

Carr, A. C. See *Goldstein, A. P.*

Cartwright, D. S. Note on "Changes in Psychoneurotic Patients with and without Psychotherapy" - - - 403

Cattell, R. B. Second-Order Personality Factors in the Questionnaire Realm - - - 411

Chambers, J. L. Perceptual Judgment and Associative Learning Ability of Schizophrenics and Nonpsychotics - - - 211

Charen, S. A Reply to Blum - - - 407

Charen, S. Reliability of the Blacky Test - - - 16

Conger, J. J., Sawrey, W. L., and Krause, L. F. A Reanalysis of Beck's "Six Schizophrenias" - - - 83

Conger, J. J., Sawrey, W. L., and Krause, L. F. Reply to Dr. Stephenson - - - 90

Copple, G. E. Effective Intelligence as Measured by an Unstructured Sentence-Completion Technique - - - 357

Crandall, V. J., and Bellugi, Ursula. Rating Personal Adjustment Through an Analysis of Social Reinforcements - - - 49

JOURNAL OF CONSULTING PSYCHOLOGY

Edited By
Laurance F. Shaffer
Teachers College, Columbia University

Advisory Editors

Donald W. Fiske
University of Chicago

Ann Magaret Garner
University of Illinois

Harrison G. Gough
University of California

William A. Hunt
Northwestern University

E. Lowell Kelly
University of Michigan

Morris Krugman
Board of Education
New York, N. Y.

Bertha M. Luckey
Board of Education
Cleveland, Ohio

Fred McKinney
University of Missouri

Anne Roe
New York, N. Y.

Carl R. Rogers
University of Chicago

R. Nevitt Sanford
University of California

Edward Joseph Shoben, Jr.
Teachers College
Columbia University

George J. Wischner
University of Pittsburgh

J. R. Wittenborn
Rutgers University

Managing Editor
Lorraine Bouthilet

VOLUME 20

1956

BUREAU OF EDNL. & PSYL. RESEARCH
LIBRARY

Published Bimonthly By

The American Psychological Association, Inc.
1333 Sixteenth St., N. W.
Washington 6, D. C.

Printed At
Lancaster, Pennsylvania

Copyright © 1956 by the American Psychological Association, Inc.

Cuddy, J. M. See *Levy, N. M.*
Dahlstrom, W. G. See *Goodstein, L. D.*
Dana, R. H. Cross Validation of Objective TAT Scoring - - - - - - - - - - - - 33
Davids, A. Some Comments on Ansbacher's Note on the "Alienation Syndrome" and Adler's Concept of "Distance" - 485
Davids, A. The Influence of Ego-Involvement on Relations Between Authoritarianism and Intolerance of Ambiguity - - - - - - - - - - - - - - - - - - - 179
Deabler, H. L. See *Saucer, R. T.*
Diamond, M. D. The Ability of Schizophrenics to Modify Responses in an Interpersonal Situation - 441
Dibner, A. S. Cue-Counting: A Measure of Anxiety in Interviews - - - - - - - - - 475
Dittmann, A. T. See *Raush, H. L.*
Doppelt, J. E. Estimating the Full Scale Score on the Wechsler Adult Intelligence Scale from Scores on Four Subtests - - - - - - - - - - - - - - - - - - - 63
Downing, W. L. See *Lotsof, E. J.*
Dugan, R. D. See *Levy, L. H.*
Ellis, A. The Effectiveness of Psychotherapy with Individuals Who Have Severe Homosexual Problems - 191
Epstein, S., and Smith, R. Repression and Insight as Related to Reaction to Cartoons - - 391
Fairweather, G. W., Moran, L. J., and Morton, R. B. Efficiency of Attitudes, Fantasies, and Life History Data in Predicting Observed Behavior - - - - - - - - - - 58
Fairweather, G. W. See *Moran, L. J.*
Fisher, S. Plausibility and Depth of Interpretation - - - - - - - - - - - - - 249
Fisher, S. See *Moran, L. J.*
Fordyce, W. E. Social Desirability in the MMPI - - - - - - - - - - - - - - - 171
Frank, G. H. The Wechsler-Bellevue and Psychiatric Diagnosis: A Factor Analytic Approach - 67
Frederickson, W. K. See *Winter, W. D.*
Fricke, B. G. Response Set as a Suppressor Variable in the OIAS and MMPI - - - - 161
Gaston, C. O. See *Taulbee, E. S.*
Gehman, I. H., and Matyas, R. P. Stability of the WISC and Binet Tests - - - - - - 150
Gieseking, C. See *Williams, H. L.*
Goldstein, A. P., and Carr, A. C. The Attitudes of the Mothers of Male Catatonic and Paranoid Schizophrenics Toward Child Behavior - - - - - - - - - - - - - 190
Goodstein, L. D., and Dahlstrom, W. G. MMPI Difference Between Parents of Stuttering and Nonstuttering Children - - - - - - - - - - - - - - - - - - - 365
Gorham, D. R. Use of the Proverbs Test for Differentiating Schizophrenics from Normals 435
Grizwok, R., and Scodel, A. Some Psychological Correlates of Humor Preferences - - - 42
Guze, S. B. See *Matarazzo, J. D.*
Harway, N. I. See *Raush, H. L.*
Hays, W. L. See *Raush, H. L.*
Heathers, Louise B. See *Merrill, R. M.*
Helfand, I. Role Taking in Schizophrenia - - - - - - - - - - - - - - - - - 37
Henry, Edith M., and Rotter, J. B. Situational Influences on Rorschach Responses - - - 457
Hodges, A. Double Alternation: A Measure of Intelligence - - - - - - - - - - - 59
Horan, E. M. Word Association Frequency Tables of Mentally Retarded Children - - - 22
Jessor, R. Social Values and Psychotherapy - - - - - - - - - - - - - - - - 264
Johnson, Laverne C. Body Cathexis as a Factor in Somatic Complaints - - - - - - 145
Jones, H. G. The Evaluation of the Significance of Differences Between Scaled Scores on the WAIS: The Perpetuation of a Fallacy - - - - - - - - - - - - - - - - 319
Kagan, J., and Mussen, P. H. Dependency Themes on the TAT and Group Conformity - 29

Kendall, Barbara S., Mills, W. B., and Thale, T. Comparison of Two Methods of Electroshock in their Effect on Cognitive Functions 423
Kenny, D. T. The Influence of Social Desirability on Discrepancy Measures Between Real Self and Ideal Self 315
Knopf, I. J. Rorschach Summary Scores in Differential Diagnosis 99
Krause, L. F. See *Conger, J. J.*
Lakin, M. Certain Formal Characteristics of Human Figure Drawings by Institutionalized Aged and by Normal Children 471
Langston, R. D. See *Murphy, D. B.*
Lawton, M. P. Stimulus Structure as a Determinant of the Perceptual Response 351
Leiding, W. C. See *Norman, R. D.*
Leventhal, H. The Effects of Perceptual Training on the Rorschach *W* and *Z* Scores 93
Levin, M. L. Validation of the Lowenfeld Mosaic Test 239
Levine, M., and Meltzoff, J. Cognitive Inhibition and Rorschach Human Movement Responses 119
Levy, L. H. A Note on Research Methodology Used in Testing for Examiner Influence in Clinical Test Performance 286
Levy, L. H. The Meaning and Generality of Perceived Actual-Ideal Discrepancies 396
Levy, L. H., Brody, Janice R., and Windman, Georgia O. The Relationship Between the Inferential Potential of Rorschach and TAT Protocols 27
Levy, L. H., and Dugan, R. D. A Factorial Study of Personal Constructs 53
Levy, N. M., and Cuddy, J. M. Concept Learning in the Educationally Retarded Child of Normal Intelligence 445
Litwin, Dorothy. See *Meltzoff, J.*
Lorr, M., and Rubinstein, E. A. Personality Patterns of Neurotic Adults in Psychotherapy 257
Lotsof, E. J., and Downing, W. L. Two Measures of Anxiety 170
Lubin, A. See *Williams, H. L.*
Lundin, W. H., and Breiger, B. A Comparison of Productivity and Fantasy on the Rorschach and Projective Movement Sequences 342
McCraven, Vivian G. See *Singer, J. L.*
McGuigan, F. J. See *Calvin, A. D.*
Mann, L. The Relation of Rorschach Indices of Extratension and Introversion to a Measure of Responsiveness to the Immediate Environment 114
Maradie, L. J. The Goal-Spurt Hypothesis and the Rorschach Test 205
Mason, C. F. Preillness Intelligence of Mental Hospital Patients 297
Matarazzo, J. D., Saslow, G., and Guze, S. B. Stability of Interaction Patterns during Interviews: A Replication 267
Matarazzo, J. D. See *Matarazzo, Ruth G.*
Matarazzo, Ruth G., and Matarazzo, J. D. Anxiety Level and Pursuitmeter Performance 70
Matyas, R. P. See *Gehman, I. H.*
Mehlman, B., and Vatovec, E. A Validation Study of the Bender-Gestalt 71
Meltzoff, J., and Litwin, Dorothy. Affective Control and Rorschach Human Movement Responses 463
Meltzoff, J. See *Levine, M.*
Merrill, R. M., and Heathers, Louise B. The Relation of the MMPI to the Edwards Personality Preference Schedule on a College Counseling Center Sample 310
Mills, W. B. See *Kendall, Barbara S.*
Mirsky, A. F. See *Rosvold, H. E.*
Moran, L. J., Fairweather, G. W., Fisher, S., and Morton, R. B. Psychological Concomitants to Rate of Recovery from Tuberculosis 199

Moran, L. J., Fairweather, G. W., and Morton, R. B. Some Determinants of Successful and Unsuccessful Adaptation to Hospital Treatment of Tuberculosis - - - 125
Moran, L. J. See *Fairweather, G. W.*
Morton, R. B. See *Fairweather, G. W.*
Morton, R. B. See *Moran, L. J.*
Murphy, D. B., and Langston, R. D. A Short Form of the Wechsler-Bellevue and the Army Classification Battery as Measures of Intelligence - - - 405
Murphy, D. G. The Revalidation of Diagnostic Tests for Alcohol Addiction - - - 301
Mussen, P. H. See *Kagan, J.*
Nitsche, C. J., Robinson, J. F., and Parsons, E. T. Homosexuality and the Rorschach - - - 196
Norman, R. D., and Leiding, W. C. The Relationship Between Measures of Individual and Mass Empathy - - - 79
Parsons, E. T. See *Nitsche, C. J.*
Pascal, G. R., and Zax, M. Psychotherapeutics: Success or Failure - - - 325
Peek, R. M., and Storms, L. H. Validity of the Marsh-Hilliard-Liechti MMPI Sexual Deviation Scale in a State Hospital Population - - - 133
Peters, G. A., Jr. Emotional and Intellectual Concomitants of Advanced Chronic Alcoholism - - - 390
Phares, E. J., and Rotter, J. B. An Effect of the Situation of Psychological Testing - - - 291
Polyot, C. J. See *Berg, J.*
Psychological Test Reviews - - - 157, 237, 321, 408, 487
Purcell, K. A Note on Porteus Maze and Wechsler-Bellevue Scores as Related to Antisocial Behavior - - - 361
Purcell, K. The TAT and Antisocial Behavior - - - 449
Raush, H. L., Sperber, Z., Rigler, D., Williams, J., Harway, N. I., Bordin, E. S., Dittman, A. T., and Hays, W. L. A Dimensional Analysis of Depth Interpretation - - - 43
Reznikoff, M., and Tomblen, D. The Use of Human Figure Drawings in the Diagnosis of Organic Pathology - - - 467
Rigler, D. See *Raush, H. L.*
Roberts, Ruth. See *Sinnett, E. R.*
Robinson, H. A. See *Silverstein, A. B.*
Robinson, J. F. See *Nitsche, C. J.*
Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D., Jr., and Beck, L. H. A Continuous Performance Test of Brain Damage - - - 343
Rotter, J. B. See *Henry, Edith M.*
Rotter, J. B. See *Phares, E. J.*
Rubinstein, E. A. See *Lorr, M.*
Rubinstein, I. See *Williams, H. L.*
Sarason, I. G. The Relationship of Anxiety and "Lack of Defensiveness" to Intellectual Performance - - - 220
Sarason, I. See *Rosvold, H. E.*
Saslow, G. See *Matarazzo, J. D.*
Saucer, R. T., and Deabler, H. L. Perception of Apparent Motion in Organics and Schizophrenics - - - 385
Sawrey, W. L. See *Conger, J. J.*
Scherer, I. W. See *Winne, J. F.*
Scodel, A. See *Grizwok, R.*
Shontz, F. C. Evaluative Conceptualizations as the Basis for Clinical Judgments - - - 223
Siegman, A. W. Cognitive, Affective, and Psychopathological Correlates of the Taylor Manifest Anxiety Scale - - - 137
Siegman, A. W. The Effect of Manifest Anxiety on a Concept Formation Task, a Nondirected Learning Task, and on Timed and Untimed Intelligence Tests - - - 176

Silverstein, A. B., and Robinson, H. A. The Representation of Orthopedic Disability in Children's Figure Drawings - - - 333
Singer, J. L., Wilensky, H., and McCraven, Vivian G. Delaying Capacity, Fantasy, and Planning Ability: A Factorial Study of Some Basic Ego Functions - - - 375
Sinnett, E. R., and Roberts, Ruth. Rorschach Approach Type and the Organization of Cognitive Material - - - 109
Sipprelle, C. N., and Swensen, C. H. Relationship of Sexual Adjustment to Certain Characteristics of Human Figure Drawings - - - 197
Sisson, B. D. See *Taulbee, E. S.*
Smith, R. See *Epstein, S.*
Soyars, M. See *Calvin, A. D.*
Sperber, Z. See *Raush, H. L.*
Stanley, W. C. See *Bosquet, K. T.*
Starer, E. The Use of the Kaleidoscope as an Adjunct to the Rorschach - - - 466
Stephenson, W. A Reply to the Reanalysis of Beck's "Six Schizophrenias" - - - 88
Storms, L. H. See *Peek, R. M.*
Swensen, C. H. See *Sipprelle, C. N.*
Taft, R. Intolerance of Ambiguity and Ethnocentrism - - - 153
Taulbee, E. S., Sisson, B. D., and Gaston, C. O. Affective Ratio and 8–9–10 Per Cent on the Rorschach Test for Normals and Psychiatric Groups - - - 105
Taylor, J. W. Relationship of Success and Length in Psychotherapy - - - 332
Thale, T. See *Kendall, Barbara S.*
Tolor, A. A Comparison of the Bender-Gestalt and the Digit-Span Test as Measures of Recall - - - 305
Tomblen, D. See *Reznikoff, M.*
Trites, D. K. Evaluation of Assumptions Underlying Interpretation of Sentence Completion Tests - - - 8
Tucker, J. E. Group Psychotherapy with Chronic Psychotic Soiling Patients - - - 430
Tyrrell, Sybil. See *Calvin, Allen D.*
Vatovec, E. See *Mehlman, B.*
Webb, W. B. A Procedure for Obtaining Self-Ratings and Group Ratings - - - 233
Wetherhorn, M. Flexor-Extensor Movement on the Rorschach - - - 204
Wickes, T. A., Jr. Examiner Influence in a Testing Situation - - - 23
Wiener, M., Carpenter, B., and Carpenter, Janeth T. Determination of Defense Mechanisms for Conflict Areas from Verbal Material - - - 215
Wilensky, H. See *Singer, J. L.*
Williams, H. L., Lubin, A., Gieseking, C., and Rubinstein, I. The Relation of Brain Injury and Visual Perception to Block Design Rotation - - - 275
Williams, J. See *Raush, H. L.*
Windman, Georgia O. See *Levy, L. H.*
Winne, J. F., and Scherer, I. W. A Second Study of Psychological Changes during the First Year Following Prefrontal Lobotomy - - - 281
Winter, W. D., and Frederickson, W. K. The Short-Term Effects of Chlorpromazine on Psychiatric Patients - - - 431
Wirt, R. D. Actuarial Prediction - - - 123
Wirt, R. D. Ideational Expression of Hostile Impulses - - - 185
Wirt, R. D., and Broen, W. E., Jr. The Relation of the Children's Manifest Anxiety Scale to the Concept of Anxiety as Used in the Clinic - - - 482
Zax, M. See *Pascal, G. R.*
Zirkle, G. A. Learning in the Immediate Post-Electroshock Period - - - 399
Zweibelson, I. Test Anxiety and Intelligence Test Performance - - - 479

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

External and Internal Stimulus Factors in Rorschach Performance

James Bieri and Edward Blacker
Harvard University[1]

It is customary to conceive of the Rorschach situation as one in which the subject (*S*) may respond to a variety of stimulus elements which are available to him in the blots. The ordering of the stimuli or determinants available to *S* usually centers about the *form* quality of the blot. That is, the essential shape of the blot may constitute the only stimulus factor utilized in a given response. However, other stimulus factors may enter into the determination of the response. These include movement, color, and shading, and each may modify or supplant the form determinant. It is possible to construe these determinants as falling on a continuum of external-internal stimulus factors. Sarason (7), for example, has spoken of external and internal referents in Rorschach performance. Thus, color is an aspect of the blot that is physically present in the blot as a stimulus. In this respect it is a readily available *external* stimulus factor. On the other hand, movement represents an *internal* stimulus factor having no intrinsic representation in the blot material. It is, so to speak, a stimulus quality more proximal to *S*, one which is apparently more a function of some internal process rather than of the external stimulus.

Thus we may assume that two categories of Rorschach responses exist, those responses whose determinants represent external qualities of the stimulus material, and those responses whose determinants are more a function of *S* than of the physical stimulus material. In the former category would be placed the form, color, and shading responses, while in the latter would be found the various types of movement responses. It would seem appropriate, therefore, to analyze and establish behavioral correlates of these two classes of conceptual responses. If these two classes of responses do indeed have different behavioral correlates, then their utilization in a broader clinical and personality framework would seem warranted.

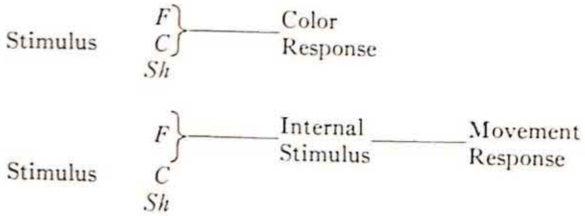
In the present study, the chief emphasis will be on the two representatives of these response categories discussed above which are utilized in Rorschach's experience-type, i.e., human movement and color responses. Starting with Rorschach himself (6), the experience-type has been a fundamental variable in Rorschach interpretation. Both human movement and color responses hold key positions because of their frequency of occurrence and of their assumed clinical significance. In addition, there is an increasing body of research literature relevant to the personality and behavioral characteristics associated with these two types of responses. Secondarily, this study will be concerned with two other kinds of responses on the external-internal stimulus continuum, i.e., shading and nonhuman movement responses, respectively. The behavioral characteristic which will be primarily employed in investigating these two classes of responses will be that of reaction time.

Investigating these two classes of responses by means of reaction time behavior can be understood when we consider the nature of the response categories discussed above. Color is an immediately available external stimulus, which may be utilized in giving a response. Movement, however, has in addition an intermediate level of response. The *S* perceives the

---

[1] This study was facilitated by a grant from the Laboratory of Social Relations, Harvard University.

form quality of the blot, apparently rejects or ignores other external stimulus factors which may be present, and invokes instead an *internal* modification of the stimulus such that movement is perceived in the blot. The intervention of this intermediate, more subjective process implies that, other things being equal, the time required for the movement response may be greater than that for the color response. Schematically, we may represent this state of affairs as follows:



Several recent studies have investigated this relationship originally discussed by Rorschach, that is, that movement or kinesthetic responses represent an absence or delay of acting-out, overt motor behavior, while color responses imply a greater tendency to immediate action. Siipola and Taylor (10), using achromatic blots, found a greater production of *M* responses under "free" conditions than under "pressure" conditions. Similar results were obtained by Singer, Meltzoff, and Goldman (11). In their study, when *S*s were forced to delay or inhibit responses, more *M* responses were produced. Other studies in this area have also obtained results suggesting a positive relationship between *M* production and inhibited motor behavior (5, 12). These studies (5, 8, 10, 11, 12) have been primarily concerned with movement responses rather than color responses. Thus, while the *M:Sum C* ratio implies a comparison of two response variables in predicting behavior, research has dealt chiefly with single variables, notably that of human movement. In the present study, primary emphasis is placed upon the comparison of experience-type patterns (introversive, extratensive, and ambiequal) rather than upon the individual components of the experience-type.

With these considerations in mind, the following problems were posed for investigation:

1. Utilizing reaction time as a behavioral criterion, are there differences in the behavior of *S*s depending upon the *kind* of experience balance produced?
2. If there is a difference in the behavior of *S*s with different experience types, is this difference specific to responses of a certain type, or is it general to all responses to the blots?
3. If behavioral differences are found in Problem 1 between *S*s who emphasize responses with internal and external stimulus factors, will the same sort of behavioral differences exist between *S*s in relation to other Rorschach responses belonging to these two categories?

Method

Owing to the nature of the study, it was necessary to modify the Rorschach procedure so that a reaction time could be obtained for each response and to insure that each *S* reacted to the identical blot material. To meet these demands, only a portion (a Rorschach *D*) of each of the standard blots was used. The portions of the blots utilized were selected according to the ease with which any given blot area could be given variable interpretations in terms of determinants and content. Thus, each blot area selected was judged capable of eliciting at least two of the four major determinants, i.e., form, color, shading, and movement. In addition, the number of achromatic and chromatic blots was in the same proportion as in the regular series (five achromatic, five chromatic blots). The blot areas chosen, in order of the administration, are listed below with Beck's number notation (1).

Card I: entire center "figure" area (*D 4*)
Card II: upper left red (*D 2*)
Card IV: lower center area (*D 1*)
Card III: upper left red inverted (*D 2*)
Card V: entire right half without "antenna" (*D 4*)
Card VIII: bottom center pink and orange area inverted (*D 2*)
Card VI: entire left half of card without top projection, card rotated 90 degrees to right (*D 4*)
Card IX: left green figure, card rotated 90 degrees to right (*D 1*)
Card VII: entire right half of card (*D 9*)
Card X: upper left blue area (*D 1*)

Each blot was prepared by making a template the same size as a Rorschach card of white posterboard in which a slot was cut large enough to expose the desired blot portion. Each template was taped to its respective card with the appropriate blot portion exposed. Thus, each blot area was exposed on a card the same size as the original Rorschach cards.

Administration

Each *S* was given the modified Rorschach individually by either of the two authors. The *S* was told that he was going to be shown a series of inkblots. He was instructed to look at the entire blot and to tell the examiner (*E*) what it looked like or what it could be. The *S* was again reminded to look at the whole blot. Following this, the first card was presented to *S*. A stop watch was held unobtrusively by *E*, and the reaction time to each response was noted. As soon as *S* gave his initial response to the card, the card was removed and replaced with the next card in the series. The *E* handed each card to *S* in the position in which it was to be viewed. If *S* started to turn the card, *E* asked him to keep it in the original position. Attempts to turn the cards were surprisingly few, and by having *S* hold each card it was possible to approximate more closely actual Rorschach testing conditions.

After *S* had given responses to the ten blots, he was then told that he would look at the blots again, and this time he was to tell *E* *what else* he saw, what else it might look like. Following this second series of the cards, the identical procedure was repeated once more. On each response of each of the three series the reaction time and content were recorded by *E*. At the end of the third series, an inquiry was conducted card by card in the usual manner. Thus, a total of thirty responses was obtained for each *S*, three associations having been given to each of the ten blots.

Owing to the nature of the modified Rorschach used, some of the *Ss* repeated a response which they had given previously to the same blot. In these cases *S* would usually say, "I can't see anything different from last time," or "It still looks like a ——— to me." Under these circumstances, *E* would record the reaction time when the statement was made, and the response would be scored as originally given.

Subjects

A total of 40 paid male undergraduate *Ss*, primarily sophomores, was used in this study.

Scoring

All the Rorschach protocols were scored by one *E*, and ten of the protocols were selected at random and scored by the other *E*. Two measures of reliability were used. First, the percentage of agreement of scoring for each protocol was obtained. The average percentage of agreement obtained for the ten records was 83. In addition, the number of responses in each determinant category was counted, and the scores for the two scorers were correlated using the rank-order coefficient. These categories included *M*, *FM*, *m*, *F*, *FC*, *CF*, and *Sh*. The *Sh* category included all shading responses of the Klopfer type as well as the use of achromatic color (*K*, *KF*, *FK*, *Fc*, *cF*, *c*, *FC'*, *C'F*, and *C'*). The *Sh* category is similar to that originally proposed by Hertz (4), except that in the present study the shading responses were not weighted but merely added to form a total *Sh* score. The average interrater reliability for these several scoring categories was .85.

Results

On the basis of the Rorschach scores, each *S* was assigned to one of three experience-type groups. The introversive (*M* > *Sum C*) group consisted of the *Ss* who showed a predominance of *M* over *Sum C*. This ratio was computed in the usual fashion with increasing weights of ½, 1 and 1½ being given to *FC*, *CF*, and *C* responses, respectively. In order to be placed into the *M* > *Sum C* group, *S* had to have an *M* score that was at least *two* higher than his *Sum C* score. A total of 17 *Ss* met this criterion and were placed in the *M* > *Sum C* group. Similarly, 11 *Ss* were placed in the extratensive (*Sum C* > *M*) group because their *Sum C* scores were at least two higher than their *M* scores. The remaining 12 *Ss* were placed in the ambiequal group inasmuch as their *M* and *Sum C* scores differed by less than two from each other.

Table 1

Mean Rorschach Scores for $M > Sum\ C$ Group ($N = 17$), $Sum\ C > M$ Group ($N = 11$) and Ambiequal Group ($N = 12$), and for All Subjects Combined ($N = 40$)

Score	$M >$ $Sum\ C$	$Sum\ C >$ M	Ambiequal	Total
M	7.9*	1.6	3.0	4.7
FM	2.8	2.4	2.2	2.5
m	1.1	1.5	0.8	1.2
F	12.5	12.8	17.3*	14.0
Sh	2.8	4.6	2.8	3.3
FC	1.5	2.1	1.5	1.6
CF	1.1	5.1*	2.4	2.6
Sum C	1.8	6.1*	3.3	3.5

* Higher than mean values of other two groups on this variable at .01 level of significance (two tails).

The mean scores for all *S*s combined and for each group are presented in Table 1. The ratio of the mean *M* and mean *Sum C* scores for the $M > Sum\ C$ group was 7.9:1.8, for the $Sum\ C > M$ group 1.6:6.1, and for the ambiequal group 3.0:3.3. The differences between the means of the *M* and *Sum C* scores for both the $M > Sum\ C$ group and the $Sum\ C > M$ group are significant below the .01 level. On the other hand, the difference between the means of the *M* and *Sum C* scores for the ambiequal group is not significant. All tests of significance between means in this study were made by the use of White's ranking procedure (3).

Table 2 presents the mean reaction time scores in seconds for each of the three groups. It will be noted from Table 2 that the mean reaction times for the $M > Sum\ C$ group for Series 2 and 3, as well as for the total reaction times, were significantly higher ($p = .05$) than the corresponding mean reaction times of the $Sum\ C > M$ group. While the ambiequal group has reaction times generally intermediate between those of the $M > Sum\ C$ group and the $Sum\ C > M$ group, these mean reaction times of the ambiequal group were not significantly different from the mean reaction times of the other two groups.

Table 2

Mean Reaction Time in Seconds for $M > Sum\ C$, $Sum\ C > M$, and Ambiequal Groups

Series	$M > Sum\ C$	$Sum\ C > M$	Ambiequal
Series 1	17.4	10.0	12.3
Series 2	29.7*	12.4	19.3
Series 3	35.4*	18.2	18.1
Total	27.4*	13.6	16.6

* Higher than corresponding values in $Sum\ C > M$ group at the .05 level of significance (two tails).

Thus, in terms of Problem 1 of the study, we may say that groups of *S*s who differ in terms of experience balance also tend to differ in terms of their reaction time behavior. These differences are generally significant between the $M > Sum\ C$ and $Sum\ C > M$ groups, the latter having shorter reaction times than the former. The ambiequal group tends to have reaction times intermediate between the predominantly movement and color producing groups.

Next we turn to Problem 2 in order to determine whether these differences are general to the *S*s' behavior on the Rorschach or whether specific kinds of responses were accountable for the differences in reaction time behavior. Specifically, we will compare the mean reaction time of each *S*'s *M* responses with his mean reaction time on his *CF* responses, the latter being selected because more *CF* responses were given by *S*s than any other type of color responses. For the 33 *S*s who gave both *M* and *CF* responses, only 16 had greater mean reaction times to *M* responses than to *CF* responses. Furthermore, if we consider the $M > Sum\ C$ and $Sum\ C > M$ groups separately, we find significant positive rank-order correlations between mean reaction time on *M* responses and mean reaction times on *CF* responses (rho = .71 and .62, respectively). These findings indicate that *S*s' reaction time behavior in this situation is a general phenomenon, and that *M* responses individually do not elicit longer reaction times than do individual *CF* responses. These findings are in agreement with other researches (10, 11) which demonstrate that the over-all response set to the Rorschach situation is related to types of responses elicited.

It should be noted that as *S*s proceeded from one series to the next, the conceptual problem of giving a response became progressively more difficult. This is reflected in the general increase in mean reaction times from series to series (Table 2). Thus, the difference between mean reaction time on series 1 and series 3 is significant at the .01 level for the *M* > *Sum C* group, at the .05 level for the *Sum C* > *M* group, but not significant for the ambiequal group. Because *S*s would at times repeat responses due to this increased difficulty, it is important to rule out the possibility that repetitions per se may have contributed to the differences found in reaction times between groups. That this is not the case is seen in the fact that there are no statistically significant differences between the mean number of repetitions among the *M* > *Sum C*, *Sum C* > *M* and ambiequal groups. These mean values are 3.8, 2.7, and 6.3, respectively. In addition, the Pearson product-moment coefficient for all forty *S*s between mean reaction time and number of repetitions is only .13.

The third question raised was whether the differences in reaction time behavior obtained in Problem 1 also obtain in regard to other determinants in the same categories as the *M* and *C* responses. Therefore, we should ask whether the differences in reaction time behavior which hold for *M* and *C* responses also hold for the other movement responses and for the shading responses we designate as *Sh*. Because of the relatively low frequency of both *FM* and *m* responses (Table 1), these two types of responses were combined into one group, namely (*FM* + *m*).

Since (*FM* + *m*) responses belong to the internal stimulus factor group, we should expect that (*FM* + *m*) responses would be positively correlated with *M* responses and reaction time, and negatively correlated with color responses and shading responses. Conversely, we expect *Sh* responses, since they belong to the external stimulus factor group, to be positively correlated with color responses and negatively correlated with *M* and (*FM* + *m*) responses and with reaction time. To analyze these relationships, product-moment correlations were computed between these variables for all forty *S*s. These results are presented in Table 3.

Three of the six variables in Table 3 had scores which were transformed by means of the logarithmic transformation in order to normalize their distributions. These were *M*, *Sum C*, and mean reaction time. The other variables had distributions which were relatively normal when original measures were utilized.

From Table 3 we see that our expectations regarding the (*FM* + *m*) and *Sh* variables are partially borne out by the correlations obtained. (*FM* + *m*) correlates .17 with *M*, .14 with mean reaction time, and − .05 with *Sum C*. None of these correlations is significantly different from zero, and these results may be attributed to chance alone. As for the *Sh* variable, we find it correlates .27 with *Sum C*, − .23 with *M*, − .50 with mean reaction time, and − .37 with (*FM* + *m*). The latter two coefficients are significant at the .01 and .05 levels respectively. The significant negative correlation between (*FM* + *m*) and *Sh* suggests that we are indeed dealing with two opposite response tendencies relative to these two variables. Thus, *S*s who give more of one will tend to give less of the other. It will be noted from Table 3 that the *Sh* variable is more negatively correlated with mean reaction time than is the *Sum C* variable. This suggests that the *M*:*Sh* ratio would be at least as efficient a predictor of reaction time as the *M*:*Sum C* ratio. Indeed, if we substitute each

Table 3

Intercorrelations of Rorschach Scores and Mean Reaction Time for All Subjects ($N = 40$)

Score	*M*	*Sum C*	*F*	(*FM* + *m*)	*Sh*	M_{RT}
M		−.42**	−.47**	.17	−.23	.35*
Sum C			−.25	−.05	.27	−.25
F				−.35*	−.38*	−.04
(*FM* + *m*)					−.37*	.14
Sh						−.50**

* .05 level of significance (two tails).
** .01 level of significance (two tails).

S's *Sh* score for his *Sum C* score in calculating his experience-type, we find that 13 of the 17 *Ss* in the $M > Sum\ C$ group and eight of the 11 in the $Sum\ C > M$ group still meet the experience-type criterion for their respective groups (χ^2 corrected for continuity, $p = .03$). Grouping *Ss* according to the *M*:*Sh* ratio in an analogous fashion to the *M*:*Sum C* groupings, we find twenty *Ss* in the $M > Sh$ group, 12 in the $Sh > M$ group, and eight in the ambiequal group. The mean total reaction times for these groups are 27.8, 11.4, and 15.4, respectively. Comparison of these values with the corresponding *M*:*Sum C* values in Table 2 shows the very close similarity between the mean total reaction times. Again, there is a significant difference ($p = .01$) between the mean total reaction time scores of the $M > Sh$ and the $Sh > M$ groups.

Discussion

The results of this study are congruent with other studies discussed earlier in indicating that the time dimension is an important variable involved in the production of Rorschach responses. This seems to be the case particularly for *M* and *Sh* responses. Although shorter reaction times are produced by the $Sum\ C > M$ group, reference to Table 3 indicates that the negative correlation between *Sum C* and mean reaction time is not significant, whereas the positive correlation between *M* and mean reaction time is significant. Thus, the presence or absence of *M* responses may be more crucial to reaction time behavior than the presence or absence of color responses. However, the significant negative relationship between *Sh* and reaction time supports the contention that short reaction time is associated with the fuller utilization of external stimulus factors in the blots whether these factors be color or shading. It may be that the side effects of color in the blots, as suggested in Siipola's work on hue-form incongruency (9), and as observed clinically in so-called "color shock," served to decrease the negative relationship between latency and the use of color.

It should be pointed out that other studies have found some indications of an inverse relationship between movement and shading responses. Cox and Sarason (2) report that while high anxiety *Ss* gave more *M*, *FM*, and *m*, low anxiety *Ss* gave more *Fc*, *cF*, and *c* responses. Similarly, Singer, Meltzoff, and Goldman (11) found a decrement in shading responses with inhibited *Ss* who produced more *M* responses. We see these findings as consistent with our thesis that different conceptual processes are involved when internal and external stimulus factors are utilized in inkblot reactions.

The ambiequal group provides several important additions to the findings. From Table 1, it will be observed that this was essentially a form response group, i.e., it produced significantly more *F* responses than the other two groups. Further, Table 2 indicates that the reaction times of this group more nearly approximated those of the $Sum\ C > M$ group than of the $M > Sum\ C$ group. Since we originally assumed form was an external stimulus factor, these findings would again be consistent with our underlying assumptions. However, several findings suggest that *F* responses may be on a different continuum than the external-internal stimulus factor continuum. Thus, if form is construed as an external stimulus factor, we should expect it to correlate negatively with mean reaction time. The findings in Table 3 show this correlation to be only $-.04$. In addition, if form is an external stimulus factor, we should expect *F* would correlate more negatively with *M* and ($FM + m$) than with *Sum C* and *Sh*. However, we find in Table 3 that *F* correlates about equally with *Sh* ($r = -.38$) and with ($FM + m$) ($r = -.35$). Thus, form may represent a separate stimulus factor which can be combined with Rorschach determinants on the external-internal continuum.

Intelligence appears to have played little role in producing the results of this study. There were no significant differences between the three groups on the vocabulary subscale of the Wechsler-Bellevue. None of the Rorschach variables studied correlated significantly with intelligence as measured by vocabulary. In addition, this intelligence measure correlated only .16 with mean reaction time.

The approach presented in this paper has emphasized the formal, conceptual properties of inkblot reactions rather than the clinical or personality correlates of such reactions. It is our belief that both these approaches are of value in helping to formulate general principles of behavior from which predictions useful to the clinician can be made. With more understanding of the conceptual processes involved in reactions to blots, and of the behaviors associated with these processes, we can proceed to develop a theoretical framework for Rorschach behavior which is both experimentally verifiable and clinically useful.

Summary

The assumption is made that all Rorschach responses can be conceptualized into two categories, i.e., those responses whose determinants include only external stimulus factors (form, color, and shading) and those responses whose determinants include internal stimulus factors (movement). The primary purpose of the study was to determine if there was a difference in the behavior of *S*s who showed a predominance of one or the other of these two response modes. Due to the conceptual differences implied by these two types of responses, it was decided that reaction time behavior might provide a critical behavioral differentiation. In order to insure that each *S* had the opportunity to give as many different kinds of responses as possible, and in order to obtain reaction times for each response, a modification of the Rorschach was used. This modification consisted of ten selected details, one from each blot, which were shown to *S* in three consecutive series, always in the same order. Consequently, thirty responses were obtained from each of the forty *S*s. Each *S* was placed in one of three groups depending upon whether his experience-type was introversive ($M > Sum\ C$), extratensive ($Sum\ C > M$), or ambiequal. The main findings were:

1. The *S*s in the $M > Sum\ C$ group generally had significantly longer reaction times to the blots than did *S*s in the $Sum\ C > M$ group.

2. This difference was general to the *S*s' performance and was not a specific effect of longer reaction times in giving movement responses or shorter reaction times in giving color responses.

3. Other responses in the internal and external stimulus factor categories were found to differentiate reaction time behavior in an analogous fashion. Thus, ($FM + m$) responses correlated positively but insignificantly with reaction time. Shading responses (Sh) correlated significantly in a negative direction with reaction time. Both ($FM + m$) and Sh responses correlated in the same direction with other Rorschach variables as did the M and $Sum\ C$ responses. The $M:Sh$ ratio proved to differentiate *S*s' total reaction time behavior as well as the $M:Sum\ C$ ratio. These findings suggest, in conclusion, that the categorization of Rorschach determinants into internal and external stimulus factors is of value in understanding the underlying conceptual processes involved in inkblot reactions.

Received June 14, 1955.

References

1. Beck, S. J. *Rorschach's test.* Vol. I. (2nd Ed.) New York: Grune & Stratton, 1949.
2. Cox, N., & Sarason, S. B. Test anxiety and Rorschach performance. *J. abnorm. soc. Psychol.,* 1954, **49**, 371–377.
3. Edwards, A. L. *Statistical methods for the behavioral sciences.* New York: Rinehart, 1954.
4. Hertz, Marguerite R. Suicidal configurations in Rorschach records. *J. proj. Tech.,* 1948, **12**, 3–58.
5. Meltzoff, J., Singer, J. L., & Korchin, S. J. Motor inhibition and Rorschach movement responses: A test of the sensory-tonic theory. *J. Pers.,* 1953, **21**, 400–410.
6. Rorschach, H. *Psychodiagnostics.* Bern: Huber, 1942.
7. Sarason, S. B. *The clinical interaction: With special reference to the Rorschach.* New York: Harper, 1954.
8. Schumer, F. C. Some behavioral correlates of Rorschach human movement responses. Unpublished doctor's dissertation, Yale Univer., 1949.
9. Siipola, Elsa M. The influence of color on reactions to ink blots. *J. Pers.,* 1950, **18**, 358–382.
10. Siipola, Elsa M., & Taylor, V. Reactions to ink blots under free and pressure conditions. *J. Pers.,* 1952, **21**, 22–47.
11. Singer, J. L., Meltzoff, J., & Goldman, G. D. Rorschach movement responses following motor inhibition and hyperactivity. *J. consult. Psychol.,* 1952, **16**, 359–364.
12. Singer, J. L., & Spohn, H. E. Some behavioral correlates of Rorschach's experience-type. *J. consult. Psychol.,* 1954, **18**, 1–9.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Evaluation of Assumptions Underlying Interpretation of Sentence Completion Tests[1]

David K. Trites
School of Aviation Medicine, USAF, Randolph Field, Texas

The present study was undertaken to evaluate two of the assumptions sometimes made in clinical, but more frequently in psychometric, or objective, use of projective-type sentence completion tests. (*a*) Incomplete sentence stimulus items whose verbal content and meaning are immediately apparent and generally agreed upon will elicit responses which refer to this generally agreed upon meaning. (*b*) Systems of response classification which represent the attitudes and objects referred to by both stimulus and response may be developed by considering only the verbal content of the stimuli.

The evaluation of these assumptions was based upon a factor analysis of the tetrachoric intercorrelations of 74 of the items from an 88 item sentence completion test which had been administered as part of an experimental battery of adaptability screening tests to aviation cadets and student officers as they entered flight training. For the correlations, the responses of 392 cadets were classified dichotomously as indicating either a positive or negative attitude with reference to adjustment to the pilot training program. Independently, responses were classified into one of 13 scoring categories covering what seemed to be the psychologically relevant areas with respect to both stimulus items and adaptability criteria.

Application of the complete centroid method of factoring yielded four factors. Blind rotation produced a relatively well-defined simple structure.

Factor I, called the *Self-Centered Anxiety Factor,* was defined by items containing a negative statement and items referring to bodily characteristics. Apparently the factor reflects a dimension not explicitly recognized when the scoring key was formulated.

Factor II, named the *Air Force Motivation Factor,* had highest loadings on items eliciting responses referring to Air Force activities, and corresponded to one of the categories of the scoring key.

Factor III, the *Interpersonal Factor,* was best defined by items representing an orientation toward interpersonal objects and pleasures. The factor seemed to be a complex of two of the scoring categories representing interpersonal activities.

Factor IV, called the *Narcissistic* (or *Self-Enhancement*) *Factor,* had its highest loadings on items referring to parents, childhood, and most of the remaining items containing a personal pronoun which were not loaded on Factor I. Some elements of one of the most complex scoring categories were described by this factor.

The extraction and interpretation of these four factors support the first assumption by defining a communality of interpretation of the item stimuli. The overlap of the four factors with the categories of the scoring key is partial support for the second assumption.

Brief Report.
Received October 25, 1955.

---

[1] An extended report of this study was published as Psychiatric screening of flying personnel: Evaluation of assumptions underlying interpretation of sentence completion tests, by David K. Trites. *Air University, USAF School of Aviation Medicine,* No. 55-33. Randolph Field, Texas, March, 1955. It may be obtained without charge from David K. Trites, School of Aviation Medicine, USAF, Randolph Field, Texas.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

The Influence of Color on Reactions to Incomplete Figures

Jacob Berg[1] and C. J. Polyot
University of Maine

Throughout the history of the Rorschach test, no problem has received more attention than the clinically observed disturbances which some subjects display when viewing the chromatic inkblots (2, 3, 7). One of the more significant experimental efforts which deals with the influence of color on reactions to inkblots is by Siipola (9) in which she disputed Rorschach's assumption that color per se endowed an inkblot with "magic, affect-arousing properties." Siipola postulated that affective phenomena such as associative blocking, strong emotional reactions, and symptoms of conceptual and behavioral disorganization which occurred in some *S*s when viewing a chromatic inkblot were probably due to the conceptual difficulty arising from "hue-form incongruity."

This difficulty is said to occur when the form of the configuration suggests a concept to which the color is not appropriate. Primarily, this incongruity creates a delay in response as indicated by the increased reaction time in responding to the chromatic blots as compared to an identical but achromatic set. At times, this incongruity also leads to significant changes in the content of responses. Since Siipola's study has had considerable influence on the theoretical re-evaluation of the role of color on reactions to inkblots (1, 2, 4, 5, 7, 8), her findings are worth considering.

Siipola cut out 20 colored portions of the original Rorschach test cards and photographed them. The achromatic set was shown to 60 women while the chromatic set was shown to 72 women. Only the first conceptual response elicited spontaneously was used. The mean reaction time between the two groups for each card was compared. She found the chromatic cards required a longer mean reaction time than the achromatic set; only two exceptions occurred where the reverse was true. Unfortunately, no statistical test was computed to determine whether these differences might not be due to chance fluctuations. Since the hypothesis of hue-form incongruity leaned so heavily on this finding, mean latency of response scores—which are notoriously skewed—hardly suffices as an adequate support. In a recent study by Lazarus and Oldfield (5) which duplicated much of Siipola's study, there was no clear-cut indication that the chromatic cards delayed the response significantly more than the achromatic mates.

A question also may be raised regarding an interpretation Siipola placed on her data when the chromatic blots were found to be either more "pleasant" or "unpleasant" than the achromatic ones. She argued that such results indicated that colored blots were more "affect-laden" in the sense that they were more likely to arouse emotional attitudes *during* the process of responding (9, p. 366). Phillips and Smith (7) raised the question: if associative blocking and other disturbing emotional effects resulted from hue-form incongruity, how could *S*s ever find such a process "pleasant"? Another interpretation is possible; namely, that these affective states depended on whether the *S*s were able or unable to give satisfactory responses to the cards. For example, with two cards equally saturated with a red hue, Siipola could not understand why one of these cards was found "pleasant," the other "unpleasant." Upon closer examination of her data, it was found that the "unpleasant" card elicited four times as many rejected or deteriorated responses as did the "pleasant" card. Rather than utilize emotional attitudes during the process of responding to explain these findings, it appears more likely that these attitudes resulted from success or failure to respond. Indeed, Lazarus and Oldfield (5) found precisely this to be true.

---

[1] Senior author is now associated with the Student Counseling Bureau, University of Minnesota.

One of the difficulties Siipola encountered was to explain why hue-form incongruity at times did not effect any differences in the content of the responses for the two groups. In resolving this problem, she was led to postulate that when a card is "highly structured," color exerts no influence since the form is too compelling. While this reasoning may be accused of being circular, some of her data contradict this notion in that so-called highly structured cards had unusually long reaction time scores. Why should this be so? To argue that "*Ss* spent considerably more time searching in vain for a more congruent concept" at the same time you argue that highly structured forms are more compelling is to eat your pie and have it too. It is difficult to measure card ambiguity using Rorschach inkblots. Lazarus and Oldfield (5) tried to establish a criterion of nonambiguity by selecting those cards, regardless of whether they happened to be in a chromatic or achromatic form, which elicited the same content in at least 77 per cent of the total responses obtained. We are not informed as to how many of the 36 cards met this criterion, nor why this cut-off point was chosen.

In the present investigation, stimulus configurations were used which allowed a more varied yet thoroughly controlled use of hue, and at the same time provided for a clear determination of the degree of ambiguity of each card. The following hypotheses were then tested:

1. The presence of hue in a configuration generally produces a significant delay in eliciting the first conceptual response.
2. This delay in response is due to the incongruity that exists between the form and the hue of the figure perceived.
3. Hue has a weak selective influence on the content of the response when the configurations are highly structured or nonambiguous; but a strong, disruptive influence when the figures are poorly structured or ambiguous.

Materials and Procedure

The configurations used were taken from the Closure Test [2] developed by C. M. Mooney of McGill University. This test is similar to the Street-Gestalt Test to the extent that the figures are incompletely drawn in black and white, and *Ss* are required to perceive the object depicted in the configuration under reduced cues. Figure 1 contains the 12 cards selected from Mooney's test for the present study. These cards were chosen because norms indicated they represented varying degrees of ambiguity or structuredness. The index of difficulty was based on the proportion of *Ss* able to identify each figure correctly in the total configuration in the absence of color. If more than 50 per cent of the *Ss* were able to do this, the figure was called *highly structured;* otherwise, it was called *ambiguous.* (The index of difficulty for each card appears in Table 2.)

Five sets of 12 figures each were used. Each set differed from the other sets in the hue applied to the white areas in each figure. Hues were chosen to represent the popular intervals of the color spectrum, and were red, green, yellow, and purple. The fifth set was achromatic. These 5 sets were identical in every respect except for the variation in hues from set to set. Each configuration measured 4 inches by 6 inches and was mounted on a white cardboard background 7 inches by 10 inches.

Two hundred and fifty undergraduates of normal color vision, enrolled at the University of Maine, were the subjects. Approximately half were women. The *Ss* were randomly assigned to one of the five groups. Each *S* saw all 12 cards of a given hue. In this way, 50 *Ss* served in each of the 5 groups. Cards were presented in the upright position, one at a time, to each *S*. A sweep-second stop watch was started at the moment of card

---

[2] This test may be obtained for experimental purposes only from the Department of Psychology, McGill University. We wish to express our appreciation to this institution for providing us with several copies of this test.

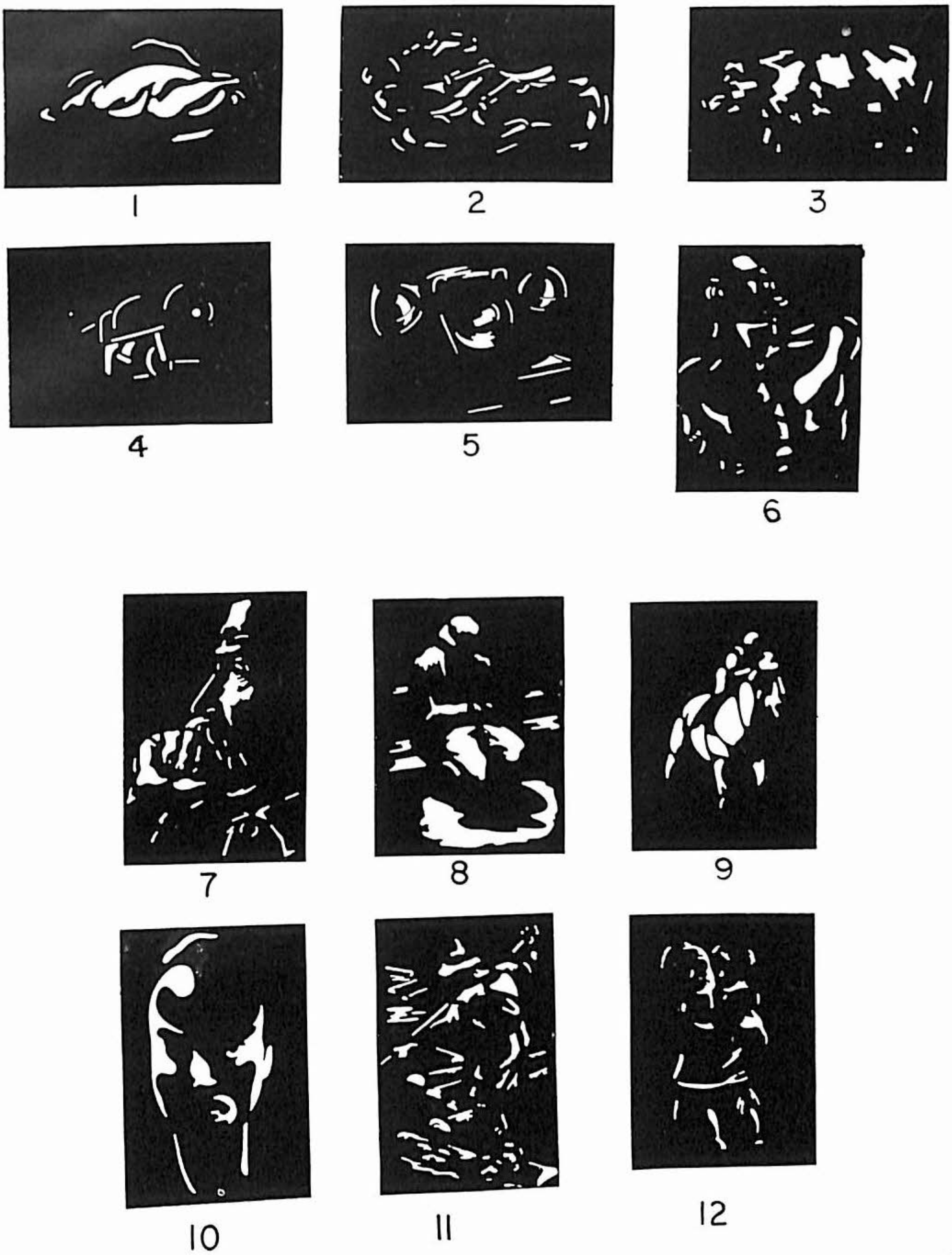


Fig. 1. The stimulus figures used in the experiment. They are (*1*) turkey on a plate, (*2*) three shoes, (*3*) pig or cow, (*4*) locomotive, (*5*) telephone, (*6*) male torso, (*7*) profile of man's face with hand on chin, (*8*) figure seated on bench, (*9*) jockey on horse, (*10*) full face, (*11*) logger with pole, (*12*) couple or person holding child.

presentation. It was stopped when *S* made a definite, scorable response. The card was then removed and the next one presented until all 12 cards had been seen. This procedure was used with all *S*s. Instructions were:

You will be shown 12 incomplete pictures of common, everyday subjects or objects one at a time. They may be confusing when first seen because some of the lines or shapes have been left out. However, as you go on looking at each card you should expect to see some popular object, figure, or subject matter. Now, some of these pictures are

Table 1

Percentage of *S*s in Each Group Offering Some Response At or Above the Median Reaction Time Score

Cards	Md (sec.)	Per cent at/above Md R.T.					Chi Square	*p*
		Ach.	*P*	*R*	*Y*	*G*		
Ambiguous								
1. Turkey	20.5	46	36	58	62	48	8.48	.05–.10
2. Shoes	15.5	52	50	42	44	62	4.94	.20–.30
3. Pig	20.5	42	46	54	52	56	2.72	.50–.70
6. Torso	20.0	40	46	64	52	50	6.31	.10–.20
7. Profile	12.8	66	42	42	62	46	10.62	.02–.05
9. Horse & Rider	12.4	48	48	54	42	60	3.74	.30–.50
Highly Structured								
4. Train	8.0	48	48	50	52	56	1.89	.70–.80
5. Telephone	6.3	46	52	36	50	64	8.22	.05–.10
8. Seated	14.1	46	50	56	46	36	4.30	.30–.50
10. Full Face	6.1	50	48	46	50	58	2.10	.70–.80
11. Logger	13.4	46	62	54	52	46	3.54	.30–.50
12. Couple	7.0	42	60	42	50	62	7.31	.10–.20

more difficult than others, and it's possible you may fail to see what is depicted. First impressions are probably best, so as soon as you recognize it, call it out so I may record your response. Have you any questions?

A time limit of 30 seconds was allowed for each card. If *S* failed to respond in this time interval, the card was taken up and the next one presented. No comment was made as to the correctness of the response.

Results and Discussion

Reaction time. A statistical problem usually arises when one attempts to deal with latency of response data. In the first place, such data rarely are normally distributed; in the second place, there are always *S*s who fail to make a response to some of the cards. The statistical technique uniquely suitable for this type of data is the nonparametric Median Test (6). Using this statistic to determine if hue alone influenced the response when content is ignored, it was found that only Card 7 differentiated the groups significantly (chi square $= 10.62$, $p = .02 - .05$). All other cards failed to do so, as Table 1 indicates. Further analysis revealed that the yellow and achromatic cards were found to be significantly different from purple, red, and green, in that the latter hues had the *shorter* latencies. After determining the median response time score for each card in each group, a Median Test was computed for all 12 cards as a group to determine whether hue had any over-all effect. Chi square was found to be 6.44, $df = 4$, $p = .10 - .20$. Similar treatment was accorded the cards when divided into the categories of *ambiguous* and *highly structured*, but no significant color influence could be established statistically.

Since the above analyses were concerned with latency of response scores regardless of the conceptual response elicited, it was important to analyze these latencies when the same correct conceptual response was given by the various *S*s in the 5 groups. Such an analysis would control still further the influence of color either as a disruptive or a facilitative force in bringing about a given response to the stimulus figures. Again, only Card 7 yielded a significant chi square at the .01 level of confidence. It was found that here the achromatic card had a significantly *longer* latency than either purple or red.

From these results, it is apparent that stimulus hue generally did not delay responses as the first hypothesis postulated. In the two instances where significant differences were obtained, either hue facilitated the response, or two different hues had differentially significant effects. Since Card 7 is essentially an ambiguous card, the second and third hypotheses are not supported.

Content of response. Siipola contended "if a stimulus form is highly structured, color will normally have no effect upon the content of the conceptual responses regardless of whether the hue happens to be congruent or incongruent with the form-favored responses" (9, p. 374). On the other hand, if a stimulus form should be ambiguous, color could produce novel concepts which otherwise were absent in the achromatic norm. Accordingly, the prediction follows that no differences should occur among the groups in their ability to recognize the figure depicted when the forms are highly structured; whereas, a difference should be manifested in the number of correct identifications when the forms are ambiguous. Since the achromatic cards were "less constraining on the associative process," they should elicit a greater proportion of correct identifications than the chromatic cards.

In Table 2, the data regarding the percentage of correct responses obtained for each card among the various groups are presented. A chi-square test was computed for each card separately, and the corresponding *p* values indicate the levels of significance attained. Five cards yielded chi-square values below the .05 level of confidence. They were cards 2 (shoes), 5 (telephone), 7 (profile), 11 (logger), and 12 (couple). Two of these cards were ambiguous, while three were highly structured. Apparently, degree of structure had little to do with this determination. Purple shoes were identified more often than yellow or red shoes; a profile in red was recognized more frequently than one in yellow; a red telephone was easier to see than a green one; a red logger was more difficult to see than a yellow or achromatic one; the couple was seen more clearly in red or achromatic than in purple. No particular color played a consistently superior or inferior role.

When the cards were analyzed as groups based on degree of structure, significant differences were not found either for the highly structured cards or the ambiguous cards. These findings do not support Siipola's hypothesis regarding the effects of color on highly structured or ambiguous cards.

In Rorschach practice, the clinician always is concerned with determining the relative influences of form and color in eliciting every conceptual response. In analyzing the data in this study for content, the responses obtained for each card by all *S*s were listed and compared. In substance, despite the variations in color used, rarely did any color differentially influence the content of the response. While there was a variety of content elicited, the kinds of content found in the achromatic cards generally turned up in all the chromatic groups. In only one single instance did a

Table 2

Percentage of *S*s in Each Group Who Were Able to Perceive the Correct Figure in Each Card

Cards	Per cent correct responses					Chi Square	*p*
	Ach.	*P*	*R*	*Y*	*G*		
Ambiguous							
1. Turkey	22	10	20	12	10	6.01	.10–.20
2. Shoes	26	36	16	14	20	10.98	.02–.05
3. Pig	42	44	28	28	38	5.04	.20–.30
6. Torso	50	40	36	40	52	4.00	.30–.50
7. Profile	32	40	48	20	38	9.48	.05
9. Horse & Rider	34	42	38	50	30	4.97	.20–.30
Highly Structured							
4. Train	58	36	46	54	46	5.80	.20–.30
5. Telephone	60	50	70	50	46	18.61	.001
8. Seated	56	54	38	50	36	6.86	.10–.20
10. Full Face	80	72	82	76	74	1.92	.70–.80
11. Logger	66	48	40	66	50	9.44	.05
12. Couple	86	68	94	78	74	13.01	.02–.03

Table 3

The Percentage of *S*s in Each Group who Failed to Offer Any Response to Each Card

Cards	Per cent rejected responses					Chi Square	*p*
	Ach.	*P*	*R*	*Y*	*G*		
Ambiguous							
1. Turkey	30	20	42	36	30	6.18	.10–.20
2. Shoes	24	16	26	24	32	3.55	.30–.50
3. Pig	36	28	46	44	46	5.17	.20–.30
6. Torso	24	36	44	34	24	6.64	.10–.20
7. Profile	34	18	16	30	22	5.28	.20–.30
9. Horse & Rider	32	16	32	30	32	12.23	.01–.02
Highly Structured							
4. Train	16	14	14	14	14	0.13	.98–.99
5. Telephone	14	12	10	24	20	5.06	.20–.30
8. Seated	22	22	36	30	36	4.76	.30–.50
10. Full Face	14	18	16	8	20	3.29	.50–.70
11. Logger	16	18	26	24	20	1.87	.70–.80
12. Couple	6	6	4	8	14	4.84	.30–.50

given concept appear with significantly greater frequency in the chromatic version. On Card 1, when seen in green, 16 *S*s called it a "leaf" compared to 4 *S*s in each of the other groups who saw it as a "leaf." This was the only instance where hue pulled a conceptual response in its direction.

It is important to note that while the materials used in this experiment differed from Siipola's blots, and the *S*s were told that a definite figure could be seen on each card, the essential task confronting *S*s in both experiments was equivalent. While Siipola instructed her *S*s that "this is a test of your imagination," she also instructed them ". . . I would like to have you tell me exactly where and how you imagined what you did . . ." It is a mistake to think that the Rorschach test is a fantasy production in the sense that the reality features of the inkblots offer no constraining rein on what *S*s see. Even though the materials used in the present experiment are not inkblots, it was found that on the highly structured cards as many as 9 different concepts were given by *S*s, while on the ambiguous cards the average for each card was 11 different concepts.

In the Siipola study, it was reported that the chromatic blots were rejected twice as often as the achromatic blots. In Table 3, no statistically significant differences among the groups as to the number of rejections elicited was found for the highly structured cards. With the ambiguous cards, only Card 9 (horse and rider) differentiated the groups at the .01–.02 level of confidence. The purple hue contributed fewer rejections than the other hues, which in turn did not differ from each other. These data indicate that rejections do not necessarily occur more frequently because hue happens to be present.

In view of the results of this experiment, it is the writers' opinion that with normally functioning *S*s, form factors are far more significant in determining what is seen in a configuration than hue.

Summary and Conclusions

This study tested three hypotheses developed by Siipola to explain the alleged disturbance observed in some *S*s clinically when viewing the chromatic Rorschach inkblots. These hypotheses are:

1. The presence of hue in a configuration generally produced a significant delay in eliciting the first conceptual response.

2. This delay is due to the incongruity which sometimes exists between the form and the hue of the figure.

3. Hue has a weak, selective influence on the content of the response in highly struc-

tured figures, but a strong disruptive one in ambiguous figures.

Incomplete figures taken from Mooney's Closure Test were used. This created the advantage of using figures which varied in degree of structure or ambiguity and allowed for a more varied and controlled use of hue which could be compared with the achromatic version. Five sets of 12 figures each were shown individually to 250 *S*s randomly assigned to one of the 5 groups. Since all sets were identical except for the variation in the hue of the figures, color influences as an independent variable could be measured.

Analyses of the data led to the rejection of all three hypotheses of Siipola. No significant differences were found among the groups on latency of the first conceptual response elicited except for one card where the achromatic form had the longer latency as compared with red and purple. Hue-form incongruity could not explain these differences. Hue had no significant influence on the content of the response as a function of the degree of card ambiguity. Also, the chromatic cards were not rejected significantly in larger numbers than were the achromatic cards.

In brief, the findings of this investigation contradict the conclusions reached by Siipola to explain the clinically observed disturbances of some *S*s when viewing the chromatic Rorschach inkblots.

Received May 5, 1955.

References

1. Benton, A. L. Experimental validation of the Rorschach: II. Significance of Rorschach color responses. *Amer. J. Orthopsychiat.*, 1952, **22**, 755–763.
2. Fortier, R. H. The response to color and ego function. *Psychol. Bull.*, 1953, **50**, 41–63.
3. Hertz, Marguerite R. Current problems in Rorschach theory and technique. *J. proj. Tech.*, 1951, **15**, 307–338.
4. Klopfer, B., *et al. Developments in the Rorschach technique.* Yonkers, N. Y.: World Book Co., 1954.
5. Lazarus, R. S., & Oldfield, Margaret. Rorschach responses and the influence of color. *J. Pers.*, 1954, **23**, 356–372.
6. Mood, A. McF. *Introduction to the theory of statistics.* New York: McGraw-Hill, 1950.
7. Norman, R. D., & Scott, W. A. Color and affect: a review and semantic evaluation. *J. gen. Psychol.*, 1952, **46**, 185–225.
8. Phillips, L., & Smith, J. G. *Rorschach interpretation: advanced technique.* New York: Grune and Stratton, 1953.
9. Siipola, Elsa M. The influence of color on reactions to ink blots. *J. Pers.*, 1950, **18**, 358–382.

Bureau Ednl. Psy. Research
DAVID H [illegible] NING COLLEGE
Dated
Acc. No

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Reliability of the Blacky Test[1]

Sol Charen
Catholic University[2]

In a recent investigation of the concept of regression (2) use of the multiple-choice questions in the Blacky test by Blum gave equivocal results. The research involved the test-retest of tuberculous patients at a period relatively early in their illness and again four months later in the final stage of recovery. The Blacky was the only instrument, in a battery of fifteen paper-and-pencil tests and the Rorschach, which gave a conflicting picture with patients either showing no evidence on it of improvement or definite evidence of regression on recovery.

Either the Blacky questions tapped personality levels not reached by the other tests or did not have good reliability. The first assumption can be challenged by results of the Rorschach which showed only minor group changes when patients recovered. Even if Rorschach and Blacky tests are assumed not to tap the same levels of personality the regression in subjects which was shown by the latter test should have been reflected by changes in defenses or in ego structure and these changes would be picked up by the Rorschach.

Measurement of reliability of the Blacky was possible in view of the test-retest situation and evidence from the other tests that no basic personality changes occurred during the four month period of recovery. Scoring of the multiple questions accompanying each Blacky cartoon was by Blum's criteria (1) with each subject scored "weak" or "strong" in various psychosexual areas. With this dichotomy of results a point distribution can be assumed and therefore r_p was the correlation coefficient used to determine reliability (3, p. 179).

The highest r_p obtained was .519 for Castration Anxiety (Picture IV) with the remaining ones low or negative. Allowing for the relatively lower values of r_p the test-retest correlations are still low enough to suggest that one reason for the equivocal results obtained with the Blacky multiple-choice questions in this study of regression in tuberculous patients might be due to the factor of low reliability of this test.

Brief Report.
Received October 17, 1955.

References

1. Blum, G. S. Revised scoring system for research use of the Blacky pictures. Male form. Ann Arbor, Mich.: Author (mimeographed), 1951.
2. Charen, S. Regressive behavior changes in the tuberculous patient. Unpublished doctor's dissertation, Catholic Univer. of America, 1955.
3. McNemar, Q. *Psychological statistics.* New York: Wiley, 1949.

---

[1] An extended report of this study may be obtained without charge from Sol Charen, 200 Rhode Island Avenue, N.E., Washington 2, D. C., or for a fee from the American Documentation Institute. Order Document No. 4740 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

[2] Now at the Montgomery County Mental Hygiene Clinic, Rockville, Maryland.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Discriminative Powers of Rorschach Determinants in Children Referred to a Child Guidance Clinic[1]

Kennison T. Bosquet
Providence Child Guidance Clinic and Brown University

and Walter C. Stanley
Brown University

The present study sought to answer two questions: (*a*) What is the relationship between the main Rorschach determinants and age in children, aged 7 through 13, who have been referred to a child guidance clinic for treatment? (*b*) How well do the Rorschach determinants discriminate between such referred children and the "normal" children studied by Ames *et al.* (1)?

Though normative in character, our first question stemmed also from certain conventions which are used in the interpretation of children's Rorschach records. These conventions consist of statements about changes in determinants as a function of age in normal children. Yet the literature contains relatively few normative studies, and those which have been published tend to be inadequate, particularly for ages 7 through 13. Thus, the present study also raised a theoretical question: Do changes in the determinants, *assumed* to take place in normal development, occur in children referred to a clinic? These hypothesized changes may be stated as follows: (*a*) there should be an increase in frequency with age in M, FC, $F+\%$, P, and R; and (*b*) there should be a decrease in frequency with age in FM, CF, C, and $A\%$.

Our interest in the second major question was also twofold. By comparing the clinic children with those studied by Ames, we would have some basis for evaluating the generality of our findings on age as well as information concerning the discriminative properties of Rorschach determinants.

Method

The Rorschach protocols of 175 boys who had been referred to the Providence Child Guidance Clinic for diagnosis and treatment were analyzed. Twenty-five Rorschach records for each of the ages 7 through 13 were selected at random from the Clinic files, containing Rorschach protocols which had been obtained and scored by the senior author as part of the routine testing program of the Clinic.

The records were scored primarily according to the Beck (2) system, with certain adaptations from Klopfer's (5) method. The main differences from Beck were the use of animal movement responses (FM), inanimate movement responses (m) and Klopfer's $F\%$, the percentage of form responses in the total number of responses. All records were analyzed for the following 15 variables: R, M, FM, m, Total M, FC, CF, C, $Sum\ C$, Shading, $H\%$, $A\%$, P, $F\%$, and $F+\%$.

Intelligence ratings were available on 172 of the 175 cases. The ratings were mostly Children's Wechsler IQ's; a few were Binet IQ's. The mean IQ was 98.55; the sigma, 17.98. The distribution was tested for skewness and found to be not significant. Thus the sample used in this study is representative of the general population in terms of tested intelligence.

One hundred fifty-three of the fathers of children in this study were alive. Chi-square

---

[1] This report is based on an M.S. thesis submitted to Brown University by the first author. The second author served as advisor.

Table 1

Medians and Quartiles of Scores on Major Rorschach Variables

Variable	%ile	Age 7	8	9	10	11	12	13
M	75	2	2	4	3	2	2	2.5
	50	1	1	2	2	1	1	1
	25	0	0	1	0	0	0	0
FM	75	2	3.5	4	4	3	6	3
	50	1	2	3	2	2	2	2
	25	0	1	1	0	1	1	1
m	75	0	1	1	0	.5	0	1
	50	0	0	0	0	0	0	0
	25	0	0	0	0	0	0	0
Total *M*	75	4	7	8	7.5	6	7	6.5
	50	2	3	5	4	3	3	4
	25	1	1	2	1.5	1.5	2	2
FC	75	.5	1	1	1	1	2	.5
	50	0	0	0	0	0	1	0
	25	0	0	0	0	0	0	0
CF	75	2.5	2	2	2	1	1.5	1
	50	1	1	1	1	0	0	1
	25	1	0	0	0	0	0	0
C	75	0	0	0	0	0	0	0
	50	0	0	0	0	0	0	0
	25	0	0	0	0	0	0	0
Sum *C*	75	3	3	3.5	2	2	2.25	2
	50	2	2	1.5	1	.5	1	1
	25	1	.75	.25	.5	0	.5	0
Shading	75	2	2	2.5	2	1.5	2	2
	50	1	1	2	1	1	1	1
	25	0	0	0	0	0	1	0
H%	75	27	25.5	25	19.5	19	19	22.5
	50	13	8	14	13	15	11	15
	25	7.5	0	8	8	7.5	7	7
*A%**	75	64.5	61.5	67.5	67.5	71.5	70	75
	50	50	52	50	62	62	60	63
	25	36	43	43.5	46	44.5	55.5	54.5
*P***	75	4	3.5	6.5	6.5	8	7	7
	50	3	3	5	5	5	5	5
	25	2	2	3	3	3.5	3	3.5
F%	75	73	65.5	68	71	74.5	70	78
	50	64	58	54	59	61	52	58
	25	46.5	38.5	32	42	54	44	52.5
*F+%***	75	75	74	84	88	89	85	88
	50	63	65	75	80	80	71	79
	25	55	46.5	62	65	60.5	67	66
R	75	17	20.5	25	23.5	21	23	22
	50	15	16	18	17	18	17	16
	25	11	11.5	13.5	13.5	13	12	10.5

* Chi-square median test p value between .02 and .01.
** Chi-square median test p value less than .01.

tests revealed that they did not differ significantly in socioeconomic status from employed males in the state of Rhode Island, as listed in the 1950 census reports (7).

Results

Since few of the distributions of Rorschach scores were normal within age groups, chi-square tests of independence were used to test null hypotheses. The over-all median of the 175 scores for each variable was used to divide each age level into two groups, one below the median and one above. A 2 × 7 table was thus obtained for each variable. The hypothesis tested for each variable was that the age groups came from a population with a common median.

Table 1 gives the median and quartiles for each age group on each Rorschach variable and summarizes the results of these chi-square comparisons. Three variables achieved statistical significance. These are *P*, *F* + %, and *A*%, all of which increased with age.

Comparison with Ames's Data

The study of Ames *et al.* (1) was made on a group consisting primarily of children of the research group which had been followed for many years at the former Yale Clinic of Child Development. This group was supplemented with children of similar intelligence and social class levels from the Connecticut cities of New Haven and Waterbury. Fifty children, 25 boys and 25 girls, were seen at each of 13 different age levels. The children were considered to be a "normal research group." Only those in the age range 7 through 10, a total of 200 children, were used for comparison with the Clinic sample of the present study, as these were the only ages in which the two groups overlapped.

The distribution of intelligence ratings in the Ames group shows three-fourths of the children to be above the "average" category, with the median and modal ratings being "superior." The Clinic children, on the other hand, show a fairly normal distribution with the median and mode within the "average" category. The two groups differ significantly in intelligence, the chi-square p value being less than .01, with the Ames group being much higher in rating.

The Ames group was a "high-level population with more than half the children in the professional group, and three-fourths in the professional and managerial groups" (1, p. 24). Ratings were based on the Minnesota Scale of Paternal Occupations (4). The Clinic children rated on the same scale show only 15 per cent in the professional and managerial groups and about three-fourths in groups comprising clerical, retail business, skilled and unskilled trades. The hypothesis that the clinic and the Ames groups were drawn from the same socioeconomic population can be rejected, the chi-square p value being less than .01.

Ames's data on Rorschach variables include both boys and girls, while data of the present study are limited to boys. However, Ames comments, "Sex differences are minimal at most of the individual ages studied and are not for the most part consistent from age to age" (1, p. 288). Accordingly, the Clinic and Ames groups were compared on the 10 Rorschach variables where scoring was comparable by means of a chi-square analysis. The 10 variables were *M, FM, FC, CF, C,* Sum *C*, *H*%, *A*%, *F*%, and *R*.

The hypothesis tested at each age level for each variable was that dividing the Clinic and Ames age groups on the basis of the Ames age group medians yields frequencies independent of the Clinic-Ames classification. The 10 variables compared at the four ages gave 40 chi squares, of which three had p values between .02 and .01, and 37 had p values of .20 or more. For the three "significant" differences, the Clinic group had more *FM* responses at age 9; a lower *F*% at age 9, and a higher *A*% at age 10. Inspection of Wilkinson's (6) tabled values for obtaining p values of .05 and .01 in a set of four independent tests reveals that these three chi squares are indicative of "real" p values of approximately .05. However, it should be mentioned that in the case of *F*% and *A*%, the direction of difference between the Clinic and the Ames age groups was not the same for the four age comparisons, thus preventing any simple interpretation of even these "corrected" p values.

A second set of chi-square tests of independence of the Clinic-Ames classifications was based on the number of children using a determinant one or more times. In this comparison only the Ames data for boys were compared with the Clinic group since the Ames data were broken down according to sex. However, the number of variables involved was smaller, as only those so listed and scored comparably by Ames could be used. The variables were M, FM, FC, CF, and C.

These five variables compared at four different age levels gave 20 chi square, of which one (that for C, Age 7) had a p value of about .02, and 19 had p values greater than .05. The chi square for C, Age 7 is not reliable as one such value in a set of four.

In a further comparison based on children using C one or more times, the four age groups were lumped and the Clinic boys were compared with the Ames boys in a single 2×2 table. This comparison yielded a chi square with a p value of about .02, the Clinic group having the higher scores.

Discussion

It is evident from the findings with the Clinic children that the 15 Rorschach variables analyzed showed a marked insensitivity to the age variable. Significant p values obtained only for P, $F+\%$, and $A\%$, each of which showed increases with age. It should be noted, however, that the first two of these variables are statistical in derivation and are in this study based on adult norms.

The third significant increase ($A\%$) is just opposite to the conventional assumption in regard to "normal" children, and might be explainable by the nature of our sample—referred children. However, when this clinic group was compared with the Ames sample, there was no significant over-all difference. This suggests that even with "normals," $A\%$ may not decrease as commonly assumed throughout the ages studied.

With regard to the nine conventional assumptions concerning the effect of age on determinants which were listed in the introduction, the data on the Clinic group affirm the applicability of two (that P and $F+\%$ should increase), negate the applicability of one (that $A\%$ should decrease), and provide neither confirmation nor negation for the remaining six conventional assumptions.

Again, one way to treat these findings would be to dismiss them on the ground that our subjects came from a Clinic population, hence were abnormal in personality development. An alternative explanation, however, is suggested by the striking similarity in Rorschach scores between the Clinic group and Ames's "normal" research group. These two groups, it will be recalled, differed markedly in intelligence, in socioeconomic status, and in the fact of referral to a child guidance clinic, yet the only indication of a real over-all difference came from one of the three analyses involving C. The alternative implication of our findings, then, is that a simple "sign" approach to the interpretation of children's Rorschach responses is relatively insensitive either to age or to differences between markedly contrasting groups of children. This conclusion is further buttressed by the fact that the mean value for $F+\%$ for 7-year-old Clinic children was 65.84. Bochner and Halpern (3) state that an $F+\%$ of 65 is to be expected in a child just reaching school age.

Summary

The Rorschach protocols of 175 referred boys, 25 at each age, 7 through 13, were analyzed for changes in 15 Rorschach determinants. Only three, P, $F+\%$, and $A\%$, showed significant changes with age, all three increasing with age. Insofar as scoring procedure was comparable, scores on the Rorschach determinants for this Clinic sample were compared with the scores reported by Ames *et al.* for a "normal research group." Although the Clinic sample was significantly poorer in intelligence and socioeconomic status, its Rorschach responding, taken as a whole, was strikingly similar to that of the Ames group. It was suggested that the major implication of the present study was that the "sign" approach to the interpretation of children's Rorschach responding is a relatively insensitive procedure.

Received May 9, 1955.

References

1. Ames, L. B., Learned, J., Metraux, R. W., & Walker, R. N. *Child Rorschach responses.* New York: Hoeber, 1952.
2. Beck, S. J. *Rorschach's test.* Vol. I. *Basic processes.* (2nd Ed.) New York: Grune & Stratton, 1949.
3. Bochner, Ruth, & Halpern, Florence. *The clinical application of the Rorschach Test.* New York: Grune & Stratton, 1945.
4. Goodenough, Florence L., & Anderson, J. E. *Experimental child study.* New York: Century, 1931.
5. Klopfer, B., & Kelley, D. *The Rorschach technique.* Yonkers, N. Y.: World Book Co., 1946.
6. Wilkinson, B. A statistical consideration in psychological research. *Psychol. Bull.,* 1951, 48, 156–158.
7. U. S. Census of Population, 1950. Report No. P-B39, Vol. II, Part 39, Ch. B, R.I. U. S. Government Printing Office, 1952.

BUREAU OF EDNL. & PSYL.
LIBRARY
1.9.70
J289

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Word Association Frequency Tables of Mentally Retarded Children[1]

Edmund M. Horan
New York, N. Y.

No adequate word association frequency tables have been available for the measurement of group contact in children. The children's tables published in 1916 by Woodrow and Lowell reported the responses of children as a group with no separation into age groups. Moreover, the method employed in their investigation was not oral but written.

The oral responses of 732 mentally retarded children, or a total of 73,200 responses, are reported in this study. The frequency tables, based on the Kent-Rosanoff stimulus set of 100 words, have been arranged by chronological age, nine through fourteen, inclusive. The children were selected from the class personnel sheets at the Bureau for Children with Retarded Mental Development in New York City. The IQ's of these children, based on individual examinations by psychologists from the Bureau of Child Guidance, ranged from 40 to 75. The median IQ was 69. The children were all native born, with the exception of 13 children born abroad. Classes for children with retarded mental development in 97 New York City public schools took part in this investigation administered during the spring of 1954.

In comparing agreement of favorite responses between children of average intelligence (Woodrow and Lowell—written responses) and adults (Kent and Rosanoff—oral responses), and between the retarded children of this study and adults, the favorite responses of the retarded children overlapped more with those of the adults. In this comparison, the tenth and twelfth years of the retarded children seemed critical because of the significant increase made in the number of their favorite responses toward the adult level.

The present investigation found a much closer similarity between children's associations and those of adults than did Woodrow and Lowell. Differences do exist but the differences are not found in the manner of associating, but, rather, in the tendency of children to offer a greater number of unique responses, to use more different words in response to a stimulus and, at times, to be unable to respond to a stimulus word.

Word meanings play a part in the formation of associations, but not as important a part as has been thought. Responses by opposites appear as favorite responses at a much younger age than has been believed. High written-word frequency does not appear to give a stimulus word any special power in word association.

A comparison of the responses of each age group of retarded children demonstrates a uniform decrease from nine through fourteen, inclusive, in the number of failures of response and unique responses, and a concurrent increase in the number of common responses, i.e., responses with a frequency of two or more.

Brief Report
Received August 2, 1955.

---

[1] An extended report of this study may be obtained without charge from Edmund M. Horan, 529 West 111 Street, New York 25, N. Y., or for a fee from the American Documentation Institute. Order Document No. 4673 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.75 for microfilm or $2.50 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Examiner Influence in a Testing Situation[1]

Thomas A. Wickes, Jr.
Purdue University

A number of investigators (1, 2, 3, 5, 6) have studied examiners and the effects examiners have on the results they elicit from examinees. These studies have been done within different contexts and from different points of view but they all point up the fact that the examiner can, in various ways, affect the results they get from examinees. Nevertheless, some examiners tend to treat certain aspects of the testing situation as if they were unimportant and thus disregard their possible effects upon gathered data. The present study was designed to see whether or not test results could be modified by some aspects of the testing situation which are sometimes treated as if they were unimportant. The experimental hypotheses tested were: (*a*) test results will be significantly modified by the perfunctory, verbal comments, "Good," "Fine," and "All right." (*b*) test results will be significantly modified by the perfunctory, nonverbal actions of smiling, nodding the head, and leaning forward in the chair.

Procedure

Thirty-six male students ranging in age from 18 to 23 years were used in the study. These students were enrolled in introductory psychology classes at Purdue University. Students were placed in one of three groups, each group consisting of twelve subjects. Hypothesis *a* was tested in one group; hypothesis *b* was tested in another group; and the third group served as a control.

Two examiners were used in the experiment. Each examiner tested six subjects in every group. The author and the second examiner both had similar training and experience in psychology and in scoring inkblot responses via Klopfer's method. Both examiners practiced voice inflections, posture, and movements so that these conditions would be as nearly equivalent as possible in the experiment.

"Test results" in the experiment were defined in terms of movement (M) responses, i.e., the inkblot induced perception of action in humans (excluding internal parts) and/or human-like action in animals or parts of animals. The test instrument was a set of 30 achromatic inkblots, devised by the author and several of his colleagues, which had been selected by means of a pilot study from an original series of 120 blots. These 30 inkblots consisted of two series of 15 blots each, numbered 1A, 2A, . . . 15A and 1B, 2B, . . . 15B.

Both A and B series of blots were composed of five blots which elicited movement responses between 45 per cent and 50 per cent of the time in the pilot study and ten blots which elicited movement responses between 20 per cent and 25 per cent of the time in the pilot study. Those five blots in each series which elicited the most movement were presented first.

In the testing situation, the subject was conducted to a small well-lighted testing room and was asked to seat himself in a chair at the right side of the desk, adjacent to, and facing the examiner. After certain preliminaries had been completed (asking the subject to withhold questions, checking him off the appointment sheet) the instructions were read to him as follows: "I am going to show you some cards with ink blots on them. I want you to tell me what one thing it looks most like to you, or reminds you of. Please look at each card carefully, turning it so that each of the four sides is toward you" At this point, the experimenter gave a demonstration in which he picked up a blot and looked at it from each of the four sides. "Take all the time you wish, then tell me the clearest thing you see. When you finish, hand me the card and I will write down your answer."

---

[1] This paper is based on a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science, to the Graduate School of Purdue University. The study was directed by Prof. A. W. Landfield.

After the subject indicated that he understood the instructions, cards 1A to 15A were presented to him, one at a time. After each response, the subject handed the card to the examiner who then recorded the response. During the presentation of the first 15 cards, the examiner remained in one position as nearly as possible, moving only his right arm to handle the cards and to record the response. Subjects in all three groups were treated in as nearly identical a manner as possible during the administration of the first half of the thirty-card set. However, during the second half of the thirty-card set (i.e., cards 1B to 15B) variation was introduced in two of the three groups.

In the first experimental group or the verbal condition group the examiner *made a comment* after each *M* response, starting with card 1B. With the first *M* response the examiner said, "Fine." On the next *M* response the examiner said, "Good." After the third *M* response the examiner said, "All right." These comments were repeated in that order after each *M* response throughout the remainder of the series. No other comment was made and the examiner controlled his body movements except for tendering the card, recording the response, and receiving each card as it was returned.

In the second experimental group or nonverbal condition group, the examiner *modified his behavior* after each response judged to be *M*, starting with card 1B. After the first *M* response, the examiner nodded his head three times. On the next *M* response the examiner smiled. With the third *M* response, the examiner leaned forward in the chair right after the response, recorded it, and returned to the initial position. These behavioral modifications were repeated in that order after each *M* response throughout the remainder of the series. No other movement was made and the examiner made no comments unless asked a direct question and an answer was unavoidable.

In the control group, all thirty cards, 1A through 15B, were administered to each subject in a manner as nearly identical as possible.

Immediately after the subject had responded to the last card, the examiner asked him to fill out a short questionnaire. The purpose of this questionnaire was to get a rough estimate of the subjects' attitudes toward the test and the experimenter. The subjects' anonymity was safeguarded in this phase of the experiment. After the subject had filled out the questionnaire, the examiner interviewed him to determine whether or not he had been aware of the nature of the experiment. It may be inferred from the interview data that none of the subjects was cognizant of the nature of the experiment, at least at the verbal level.

Results

Prior to analyzing the data, the author excluded four subjects who gave no *M* responses to the last 15 cards under the verbal and nonverbal conditions. The total number in each group was then eight in the verbal condition group, eleven in the nonverbal condition group, and twelve in the control group.

In analyzing the data, the procedure was to compare the mean number of *M* responses elicited by the first 15 cards to the mean number of *M* responses elicited by the second 15 cards. This was done for all three groups. As recommended by McNemar (4, pp. 225–

Table 1

Difference Between Correlated Means *t* Test for Verbal Condition, Nonverbal Condition, and Control Condition

Group	*N*	*SD* of distr. of diffs. between paired scores	Computed *t*	Significance level	
				$t < .025$	$t < .005$
Verbal condition	8	3.04	2.90	2.36	3.50
Nonverbal condition	11	1.25	7.47	2.23	3.17
Control condition	12	1.63	.71	2.20	3.11

Table 2

Difference Between Uncorrelated Means t Test for Examiners in Verbal Condition, Nonverbal Condition, and Control Condition

Condition and examiner	No. tested	Mean of M diff. score for each examiner	Computed t	Significance level $t<.05$
Verbal condition				
Examiner I	5	4.40	1.49	2.36
Examiner II	3	1.67		
Nonverbal condition				
Examiner I	7	2.71	.35	2.23
Examiner II	4	3.00		
Control condition				
Examiner I	6	.83	.46	2.20
Examiner II	6	1.67		

226), the test of significance used made allowance for the fact that the two sets of scores were not random with respect to each other. The appropriate statistic in this case was the t test of difference between correlated means. Table 1 gives the results of this t test.

As can be seen from Table 1, the difference between the mean number of M responses seen in the first 15 cards and the mean number of M responses seen in the second 15 cards in the verbal condition group was significant beyond the .025 level. The difference between the mean number of M responses seen in the first 15 and the last 15 cards in the nonverbal condition group was significant beyond the .005 level. There was no significant difference between the first and last halves of the cards in the control condition group.

To gauge the effect the examiners, per se, might have had in eliciting M responses from the subjects, a t test was executed between the examiners. Essentially, this was a t test between uncorrelated means such as is outlined in McNemar (4, Ch. 12). It involved testing the hypothesis that there was no significant difference between the means of the M difference scores obtained from the various subjects tested by each examiner in each group. Table 2 gives the results of this t test.

Upon examination of Table 2, it can be seen that there was no significant difference between the examiners in any of the three groups. That is, one examiner did not seem to draw more M responses from his subjects than the other examiner.

The questionnaire or opinion sheet, as indicated previously, was designed to gain only a rough estimate of the subjects' motivational involvement in the testing situation. The results of the questionnaire indicated that approximately 90 per cent of the subjects in both the experimental and the control groups had a high degree of interest in both the experimenter and the experiment. It was felt that these conditions led to a set toward, or an involvement in the situation conducive to optimal results.

Discussion

The hypothesis that test results will be influenced by what the examiner says and does more or less routinely in the testing situation seems to have been substantiated, at least in terms of this experiment. In interpreting these results, it should be kept in mind that the samples used were small, but that the differences obtained in these small samples had to be proportionately larger before the chosen significance level could be reached.

Since the assumptions underlying the statistics used are difficult to demonstrate, it does not seem entirely legitimate to assume without further study that given a different population and/or different tests, the results would be the same. However, using the present study as a predictive basis, it is indicated that examiners need to be more careful of what they say and do, in a perfunctory manner, when they are engaged in testing. At least, they should be cognizant that even under what would be assumed to be "standardized" conditions, it is possible for their behavior to be reflected in the test results.

Summary and Conclusions

This experiment was designed to test the general hypothesis that test results will be modified by those aspects of the testing situation which are sometimes not carefully controlled or are treated as if they were unimportant. Thirty-six male subjects in two experimental groups and one control group were used to study the effects of perfunctory verbal comments and nonverbal actions on test results. The findings of the study suggest that such comments as "Good" or "Fine" and such actions as smiling and nodding by examiners have a decided effect upon test results. Thus it was indicated that examiners should be alert to the fact that even under presumably "standardized" conditions, it is possible for their behavior to be reflected in test results.

Received May 13, 1955.

References

1. Gibby, R. G., Miller, D. R., & Walker, E. L. The examiner's influence on the Rorschach protocol. *J. consult. Psychol.*, 1953, **17**, 425–428.
2. Greenspoon, J. The effect of verbal and mechanical stimuli on verbal behavior. Personal communication, 1954.
3. Marine, E. L. The effect of familiarity with the examiner upon Stanford-Binet test performance. *Teach. Coll. Contr. Educ.*, Columbia Univer., 1929, No. 381.
4. McNemar, Q. *Psychological statistics.* New York: Wiley, 1949.
5. Sacks, E. L. Intelligence scores as a function of experimentally established social relationships between child and examiner. *J. abnorm. soc. Psychol.*, 1952, **47**, 354–358.
6. Schachtel, E. G. Subjective definitions of the Rorschach test situation and their effect on test performance. Contributions to an understanding of Rorschach's test, III. *Psychiatry*, 1945, **8**, 419–448.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

The Relationship Between the Inferential Potential of Rorschach and TAT Protocols

Leon H. Levy, Janice R. Brody, and Georgia O. Windman
Indiana University

A problem frequently faced by the clinician in diagnostic work is that of deciding when to stop testing and when to continue. This is particularly true in those instances where the client has produced one test protocol which might be said to be low in *inferential potential,* i.e., a protocol which does not seem to lend itself to the derivation of meaningful clinical hypotheses. The problem investigated in this article might be stated in the form of the question: Given a test protocol of a certain relative level of inferential potential by a particular *S*, to what extent might one expect a second and different test protocol to be of the same relative level? Stated somewhat more elegantly, we are interested in the extent to which the variance in inferential potential might be accounted for by the individual subject (*S*) and by the subject-test interaction.

If it were found that the individual *S* accounted for a major portion of the variance, i.e., if one could predict inferential level from one test protocol to the next for any particular *S*, then there would be some basis for deciding whether to continue or discontinue testing on the basis of tests already administered. If, for example, the *S* produced a rather sterile Rorschach, he might be labeled "defensive" and sent on his way, or it might be decided that the TAT would be the method of choice with him. If, on the other hand, it were found that the subject-test interaction accounted for the major portion of the variance then there would be no way of predicting inferential potential from one test to the next, and one would be justified in trying several tests until one was found which permitted an adequate formulation of the case.

Method

The Rorschach and TAT were the two tests used in this study. Twelve cases were pulled from the clinic files which met the following criteria: (*a*) each folder contained a Rorschach and a TAT administered by the same examiner at the same session, and (*b*) the same examiner administered the tests in all of the cases. In addition, only male college students were used. In this way it was hoped to minimize any effects of examiner differences, changes in attitudes, range of talent, etc. The cases were selected from the files by the clinic secretary who was unaware of the purpose of the study. All Rorschachs and TAT's were then pulled from the case folders and coded.

Because of the complexity of the criterion, it was decided to use the paired-comparisons method in arriving at rankings of protocols on relative levels of inferential potential. Thus all twelve Rorschachs were ranked by means of paired comparisons and then correlated by means of rank-order correlation with the twelve TAT's similarly ranked. Judges were instructed to indicate which protocol of each pair they felt would permit them to make a larger number of inferences about the personality of the person producing the protocol. In order to measure interjudge agreement three judges were used: Judge A ranked both Rorschachs and TAT's, Judge B ranked only the Rorschachs, and Judge C ranked only the TAT's.

Results

Rank-difference correlations were calculated in order to estimate interjudge agreement as well as the relationship between relative level of inferential potential on the two tests. In each case their statistical significance was tested by reference to Olds's (1) tables of the distribution of sums of squares of rank differences.

The value of rho for Judges A and B on Rorschach rankings was .94 while the rho between Judges A and C on TAT rankings was .92. Both of these are significant at the .01 level. Because of the apparently high level of agreement between judges, the ranking of Rorschachs and TAT's on relative level of inferential potential was based on the combined rankings of the two judges for each test. Rho between Rorschach and TAT rankings was .65 which is significant at the .02 level.

Discussion

The data seem to indicate that the judges were in agreement with each other in their use of the criterion of inferential potential. This of course does not imply that had they each attempted to derive inferences from the material there would necessarily have been any substantial agreement between them either in the content or the number of these inferences.

Our findings would seem to support the hypothesis, for these two tests at least, that the individual *S* accounts for a major portion of the variance in inferential potential and, consequently, that one should be able to predict relative level of inferential potential from one test to another for any given *S*.

To the extent that these findings may be generalized to other populations, their practical implications are fairly obvious, viz., that where one has obtained a Rorschach (TAT) which is fairly barren in diagnostic significance, there is little point in administering a TAT (Rorschach) in the hope that perhaps this test will do the trick. They fail to support the assumption sometimes made that the more tests we have on the individual the more we are likely to learn about him.

Theoretically, these findings raise several interesting issues. For one thing, it may well be that one of the most diagnostically significant findings in a particular case is that the *S* produces a protocol either high or low in inferential potential, i.e., we may be dealing here with an important personality variable. Research directed along these lines attempting to determine the behavioral correlates of this variable may go a long way toward enhancing our understanding of personality and reducing the clinician's feelings of inadequacy when he comes up against a client who "doesn't give."

Another issue is the economic one. The clinician faced with the task of evaluating a client usually has only a limited amount of time in which to accomplish this objective. How can this time be spent most profitably? Our findings provide no answer to this question other than a negative one. It would seem though that some criterion, undoubtedly much more elaborate than that used in this study, might be employed so that the clinician could decide when he has reached the point of diminishing returns with a client.

Summary

Judges ranked 12 Rorschach and TAT protocols on the basis of their relative inferential potential. A significant relationship was found between the two tests and this was interpreted as indicating that the individual *S* accounts for a major portion of the variance in inferential potential. Certain practical and theoretical implications of these findings were discussed.

Received May 20, 1955.

Reference

1. Olds, E. G. Distribution of sums of squares of rank differences for small numbers of individuals. *Ann. math. Statist.*, 1938, **9**, 133–149.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Dependency Themes on the TAT and Group Conformity[1]

Jerome Kagan and Paul H. Mussen
The Ohio State University

Many personality theorists use the concept of "dependence" in explanatory formulations and clinicians often make statements about the strength or intensity of an individual's dependent needs. Although the specific behavioral referents for this concept have not been explicitly or unambiguously defined, there is probably general agreement that dependent tendencies include "the need for emotional or authoritative support in most situations, difficulty in making independent decisions and taking on responsibilities and the dread of loneliness" (5, p. 392).

This statement implies that the dependent person perceives others as more competent than he, and he is likely to seek and accept the guidance and advice of others when he has to make a decision. Since the dependent individual would regard the opinion of a group as wiser and more reliable than his own, it might be predicted that he would conform with group judgments. Thus, if the tendency to conform to group opinion is accepted as one behavioral referent for dependency, clinical evaluations of this characteristic could be validated by correlating the clinical measures with group conformity behavior.

The Thematic Apperception Test is often used as a technique for evaluating the differential strengths of various motives and there have been numerous attempts to relate the needs revealed in TAT stories to overt behavior.

Apparently, the relationship between fantasy needs and corresponding overt behavior is better for some motives than for others. Murray (6), studying a group of college men, obtained correlations of over + .40 between TAT fantasies involving nurturance and passivity and overt behavior ratings of these variables but found correlations of near zero between sexual and aggressive needs and their corresponding overt behaviors.

Sanford (9) and Tompkins (10) have suggested that if the motive in question is prohibited or punished in the individual's social milieu, e.g., aggression or sex in the middle class, there is apt to be little or no relationship between the strength of the fantasy need and the corresponding overt behavior. Thus, Bach (2), Korner (4), Pittluck (8), and Sanford (9) did not find direct relationships between the amount of aggressive TAT or doll play fantasy and the strength of overt aggression among children or adults.

On the other hand, if the motive in question is culturally sanctioned, there is more likely to be a positive relationship between fantasy and overt expression. Thus, Mussen and Naylor (7) found a positive relationship between TAT fantasy aggression and overt aggressive behavior among boys of the lower class, for whom such behavior is often approved.

Dependent behavior, also, has some cultural sanction, for seeking advice or help in a problem situation is often encouraged and praised. For this reason one might predict a more direct relationship between dependent fantasy themes and overt dependent behavior. The present research was designed to test the hypothesis that there is a positive association between TAT dependency themes and the tendency to adopt objectively inaccurate group judgments in a situation where there is strong social pressure favoring conformity.

---

[1] The authors wish to acknowledge the assistance of Mr. John Anderson in the execution of this study.

Method

The subjects were 27 male undergraduates enrolled in the elementary psychology course at The Ohio State University. Their ages ranged from 17 to 31 years with a mean of 21.6.

In the first part of the experiment, each subject was instructed to write stories in response to eight TAT cards from the Murray series in a prescribed order (1, 6BM, 7BM, 3BM, 12M, 13B, 14, and 18BM). He was told to make each story about a half a page long but not to write for more than two or three minutes on each card. Previous research (3) indicates that there are no important differences between the thematic contents of written and oral TAT stories.

After he completed the TAT, the subject was asked to cooperate in a seven- or eight-minute experiment on vision, conducted by another experimenter who needed additional subjects. This second situation was identical with the conformity situation described by Asch (1). The subject was led into a small room with four other men. Unknown to the experimental subject, these other four men were paid "stooges" cooperating with the experimenter. When the five men were seated, the experimenter gave the following instructions:

> This is a task which involves the discrimination of lengths of lines. You see the pair of white cards in front. The card on the left shows a single line; the card on the right has three lines differing in length. They are numbered 1, 2, and 3 in order. One of the three lines on the card on the right is equal in length to the standard line on the card on the left. You will decide in each case which is the equal line. You will state your judgment in terms of the corresponding number, either one, two or three. There will be 12 such comparisons. Since the number of lines is few and the group small, I shall call upon each of you in turn to announce your judgment, which I shall record here on a prepared form. Please be as accurate as possible. Suppose we start at the right and proceed to the left.

The naive subject was always seated fourth from the right and was always the fourth person to call out his judgment. When all five subjects had stated aloud their judgments on one pair of cards, these two cards were removed and replaced by a new pair with new standard and comparison lines. This was repeated 12 times with 12 different pairs of cards. The measurements of the standard and comparison lines were identical with those used by Asch (1).

The four men cooperating with the experimenter had been previously instructed to announce correct answers on five trials (1, 2, 5, 8, and 11) and to announce incorrect answers on the other seven trials (3, 4, 6, 7, 9, 10, and 12). On each of these seven crucial trials, all four announced the *same* incorrect answer and constituted a majority opinion.

The TAT stories were analyzed for two types of themes suggesting dependency needs. Themes in which the hero sought help from another individual in solving a personal problem or was disturbed over the loss of a source of love and support were classified as *D*.

It should be noted that *D* themes are similar to those which Murray considers indicative of *n* Succorance. Only one *D* theme was scored for each story.

The second category, *d* themes, were scored when the hero was given some help or gift (advice, food, money, etc.) which was *not* specifically requested. This type of story, closely related to Murray's category of *p* nurturance, is probably a more indirect measure of dependency. However, some clinicians might argue that stories in which the hero receives unsolicited nurturance reflect the story teller's dependency needs.

All stories were independently scored by the authors without knowledge of the subject's group conformity behavior. Percentages of agreement for *D* and *d* themes were 93 and 95 per cent, respectively, indicating that the scoring system was highly reliable.

Results

Sixteen of the 27 subjects conformed to the incorrect group judgment on one or more of the seven crucial trials. The number of "yields" per subject (conformity with the incorrect judgment) ranged from zero to seven with a median of one for the entire group. There was no significant difference between

Table 1

Distribution of "Yielders" and "Nonyielders" Among Subjects High and Low in *D* Themes

Group	*D* Themes High	*D* Themes Low
Yielders	10	6
Nonyielders	0	11

$x^2 = 8.52$, $p < .01$.

Table 2

Distribution of "Yielders" and "Nonyielders" Among Subjects High and Low in *d* Themes

Group	*d* Themes High	*d* Themes Low
Yielders	9	7
Nonyielders	4	7

$x^2 = .39$, $p > .50$.

the mean ages of those who yielded (22.3) and those who did not (20.6).

For this group of subjects production of *D* themes was relatively infrequent, the range being zero to three. Fifteen of the 27 subjects wrote at least one *D* theme and the median value for the group was one. Twenty-one of the 27 subjects produced one or more *d* themes with a range of one to five and a median of one for the entire group.

In order to test the relationship between dependency themes and conformity with group opinion, the 27 subjects were divided into groups of "yielders" (those who conformed at least once) and "nonyielders." The distributions of *D* and *d* scores were dichotomized into high (above median) and low (median and below) groups.

Tables 1 and 2 show the frequency of the number of yielders and nonyielders among those with high and low *D* and *d* scores, respectively. The values obtained when chi-square tests, using Yates's correction, are applied to these data are also given.

Discussion

The results indicate that conformity to group opinion can be predicted from certain kinds of dependency themes on the TAT. Table 1 shows that every one of the 10 subjects who produced two or more *D* themes yielded to the incorrect majority opinion.

On the other hand, the more indirect measure of dependency, *d* themes, was not sensitive in predicting group conformity. Apparently individuals who see their heroes as actively seeking aid are more likely to behave dependently (as defined here) than those individuals who portray their heroes receiving unsolicited nurturance.

In addition to the need for support in making decisions, fear of loneliness and isolation might characterize the dependent individual. Thus, fear of alienation by the group for disagreeing with the majority judgment may have motivated conformity with the incorrect group opinion. Minimal support for this notion is found in the TAT themes in response to the picture of the boy sitting alone on the door step of a log cabin (Card 13B). Although themes of loneliness were infrequent for these subjects, four of the five individuals who described the boy as lonely, yielded to the majority opinion.

The unique qualities of the conformity situation suggest the need for caution in generalizing the implications of these findings. Although the subjects who wrote several *D* themes conformed in a pressure situation with same sex peers, these men might react differently with individuals of different sex or status, e.g., younger women, authority figures, lower class individuals, etc.

Moreover, the pressures to conform in this situation were apparently very strong, for all subjects, yielders and nonyielders, appeared uncomfortable when their judgments disagreed with the incorrect majority. Those who yielded in this laboratory situation, therefore, might not conform in social contexts characterized by considerably less group pressure.

Summary

The purpose of this study was to relate dependency themes on the TAT to the tendency to conform to group opinion.

Twenty-seven male undergraduates wrote stories to eight TAT cards and then were individually observed in the Asch conformity

situation. In this situation, the subject and four other male students were requested to discriminate lengths of lines and orally announce their judgments. Unknown to the experimental subject, the other four men were cooperating with the experimenter and on prearranged discrimination trials, announced incorrect answers in an attempt to pressure the naive subject to conform with the incorrect majority judgment.

The subjects who produced TAT themes in which the hero sought help in a problem situation or was portrayed as disturbed over loss of sources of love and support yielded to the incorrect majority more frequently than those subjects not writing these types of stories ($p < .01$).

Received June 6, 1955.

References

1. Asch, S. E. *Social psychology*. New York: Prentice-Hall, 1952.
2. Bach, G. R. Young children's play fantasies. *Psychol. Monogr.*, 1945, **59**, No. 2 (Whole No. 272).
3. Eron, L. D., & Ritter, Anne M. A comparison of two methods of administration of the Thematic Apperception Test. *J. consult. Psychol.*, 1951, **15**, 55–61.
4. Korner, Anneliese F. *Some aspects of hostility in young children*. New York: Grune & Stratton, 1949.
5. Maslow, A. H., & Mittelmann, B. *Principles of abnormal psychology*. New York: Harper, 1954.
6. Murray, H. A. *Thematic Apperception Test manual*. Cambridge, Mass.: Harvard Univer. Press, 1943.
7. Mussen, P. H., & Naylor, H. K. The relationships between overt and fantasy aggression. *J. abnorm. soc. Psychol.*, 1954, **49**, 235–240.
8. Pittluck, P. The relation between aggressive fantasy and overt behavior. Unpublished doctor's dissertation, Yale Univer., 1950.
9. Sanford, R. N., Adkins, M. M., Miller, R. B., *et al.* Physique, personality and scholarship: a cooperative study of school children. *Monogr. Soc. Res. Child Developm.*, 1943, **8**, No. 1.
10. Tomkins, S. S. The present status of the Thematic Apperception Test. *Amer. J. Orthopsychiat.*, 1949, **19**, 358–362.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Cross Validation of Objective TAT Scoring[1]

Richard H. Dana
St. Louis State Hospital

Objective TAT scoring categories have only recently been introduced (2, 6). Three of five categories originally proposed as significant for diagnostic differentiation successfully distinguished between normal and abnormal male groups (2). As prediction measures two of these categories, Perceptual Organization and Perceptual Range, identified 76 and 70 per cent of normals, 60 and 64 per cent of neurotics, and 78 and 84 per cent of psychotics, respectively in the experimental groups. Perceptual Personalization prediction scores isolated 90 per cent of the normal group. These figures are the result of use of normal-neurotic and neurotic-psychotic medians as criteria.

The application of these three scoring categories to other populations, i.e., cross validation, and the accessibility of results on increasingly larger groups is a necessary further development.

Scoring System

Three scoring categories, Perceptual Organization (*PO*), Perceptual Range (*PR*), Perceptual Personalization (*PP*), were selected from the five original scoring categories on the basis of their prediction scores.

Perceptual Organization. The category reflects the *S*'s ability to follow the standard directions to "tell a story." Components included are card description, present behavior, past events, future events, feeling, thought, and outcome. One point is scored for each component used in the story.

Perceptual Range. Three separate stimulus properties were selected for each card on the basis of utilization by at least 90 per cent of a "normal" group (7). *PR* is designed to sample the extent to which the usual or "popular" appears in the story. A list of the 15 stimulus properties follows: Card II (*a*) family: young girl; woman, activity specified; adult male; (*b*) fields or farm; (*c*) books or school; Card III (*d*) figure, sex and age specified; (*e*) emotions noted; (*f*) activity specified; Card IV (*g*) man, emotions noted, activity specified; (*h*) woman, activity specified; (*i*) conflict or cooperation; Card VI (*j*) man, emotions noted, activity specified; (*k*) woman, emotions noted; (*l*) personality referent; Card VII (*m*) older male, activity specified, relationship specified; (*n*) male, emotions noted; (*o*) personality referent. One point is scored for each of these mentioned in the TAT story. All items included in each point must be mentioned for score to be earned.

Perceptual Personalization. Some expressions, words, and phrases used in the story are incongruous and have no obvious reference to the story that *S* is trying to relate. These inclusions are clearly neither stimulus reproductions nor additions to the stimulus. *PP* are deviations from the relatively consistent, organized, coherent protocol-product, the TAT story. These deviations, in order to be scored, must be extreme. They may refer to things labeled performance adequacy, comments, parenthetical remarks, qualifications, picture criticisms, adventitious descriptions, vagueness, evasion, or direct personal reference. One point is scored for each personalized inclusion.

Problem

Cross validation is an attempt to ascertain how well these selected TAT scores will function when applied to different samples of *Ss*. If these TAT scores are to be usable on groups other than those on which they were developed, they should be significantly related to clinical diagnosis on cross-validation samples. Furthermore, all three scoring categories should significantly differentiate between normal and clinical groups.

---

[1] This study was carried out in connection with an investigation supported by a research grant from the National Institute of Mental Health, of the National Institutes of Health, Public Health Service.

Subjects. The TAT stories resultant from administration of cards 2, 3BM, 4, 6BM, 7BM (5) were collected from three groups of male *Ss*: normals, neurotics, psychotics, 30 in each group. The *Ss* were chosen in terms of the same criteria as in the original study with one difference: the neurotic *Ss* were not hospitalized but were outpatient clinic patients.

Procedure. The 450 stories obtained from the three groups were scored for each of the three categories. The five stories from each *S* were treated as a unit. These units were coded, randomized, and scoring was done following mimeographed directions on a specially prepared score sheet by the author and a clerk.[2]

Results

Reliability. Two kinds of scorer reliability were computed by means of percentage of agreement: (*a*) scoring category reliability; (*b*) reliability of the items composing the categories. The appropriateness of percentage of agreement for estimating projective test data reliability has been discussed previously by the author (3, 4) and others (8).

Scoring category reliability was determined from the results of two scorers on 75 stories, 25 from each group, chosen at random, and scored in random order to avoid knowledge of the group to which any particular record belonged. The reliabilities of *PO*, *PR*, and *PP* were 93, 90, and 75 per cent of agreement, respectively. The over-all reliability of these three scoring categories was 89 per cent. These figures are somewhat higher than those reported in the original study. These differences can be attributed to: (*a*) practice in scoring obtained by both scorers; (*b*) the formalization of scoring criteria on a compact, easily used score sheet.

The reliability of scoring separate *PO* and *PR* items ranged from 80 to 100 per cent of agreement. The percentages of agreement for *PP* were 81, 73, 74, for normals, neurotics, and psychotics, respectively.

Validity. Validity, with diagnosis as the external criterion, was obtained by use of a nonparametric median test. Table 1 presents the median and range for each of the three groups on each scoring category. The scores in each category were placed in rank order and combined medians obtained for the groups being compared: normals-neurotics, normals-psychotics, neurotics-psychotics. The number of cases falling above and below these combined medians was determined by interpolation. Chi square was then used to determine differences between groups (Table 2).

The median test results indicate: (*a*) *PO* and *PR* significantly differentiate between normal, neurotic, and psychotic groups. (*b*) *PP* significantly differentiates between normal and clinical groups, but does not distinguish between clinical groups. Thus, cross validation of *PP* does not substantiate the significant

Table 1

Median and Range for Normal, Neurotic, and Psychotic *Ss* on Perceptual Organization (*PO*), Perceptual Range (*PR*), and Perceptual Personalization (*PP*)

Category		Median	Range
PO			
	Normal	26.50	22–31
	Neurotic	18.50	11–24
	Psychotic	10.78	6–21
PR			
	Normal	13.38	11–15
	Neurotic	8.78	5–11
	Psychotic	5.17	1–9
PP			
	Normal	.29	0–6
	Neurotic	3.83	0–29
	Psychotic	5.00	0–40

---

[2] Mimeographed copies of the TAT scoring directions and the score sheet may be obtained from the author upon request, or from the American Documentation Institute. To obtain them from the latter source, order Document No. 4743 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Table 2

Summary of Median Test Results on *PO, PR, PP* Scores of Three Groups

Category	Combined median	Above median	χ^2	p
PO				
Normal Neurotic	23.00	29.0 1.0	52.3	<.001
Normal Psychotic	21.50	30.0 .5	58.0	<.001
Neurotic Psychotic	14.00	28.0 2.0	45.1	<.001
PR				
Normal Neurotic	11.07	29.8 .3	58.0	<.001
Normal Psychotic	9.50	30.0 1.0	56.1	<.001
Neurotic Psychotic	7.00	29.0 4.8	39.7	<.001
PP				
Normal Neurotic	1.95	8.2 27.1	24.6	<.001
Normal Psychotic	2.60	6.2 27.4	30.4	<.001
Neurotic Psychotic	4.33	15.0 18.0	.6	—

differentiation obtained between neurotic and psychotic groups in the original study.

An approximate prediction measure was employed to determine the extent to which individuals are identified correctly by each of these scoring categories. The two medians reported in the original study for each scoring category were used as criteria: normal-neurotic and neurotic-psychotic. The median scores used for *PO* were 24.50 and 16.94; *PR* scores were 11.19 and 7.68; and *PP* scores were 1.90 and 7.37. Scores below the normal-neurotic median were given 0; scores falling between the normal-neurotic median and the neurotic-psychotic median were given 1; and scores above the neurotic-psychotic median were given 2.

Table 3 shows the results for each scoring category. These figures may be combined in various ways to provide the discrimination of each scoring category with each cutoff point, 0, 1, 2. Maximum discrimination is found by selecting a cutoff score which exaggerates the differences between groups. Thus, *PR* remains the "best," i.e., maximally discriminating category on which 98 per cent of normal *Ss* had 0, 81 per cent of neurotic *Ss* had 1, and 84 per cent of psychotic *Ss* had 2 scores.

Comparison of these prediction scores with those obtained in the original study (2) indicates that both *PO* and *PR* retain their effectiveness on new populations. In fact, the prediction level for the cross-validation groups is somewhat higher than in the original groups. The level of prediction for *PP* is substantially equivalent to that originally obtained. *PP* remains effective for distinguishing normal from pathological groups but not for more precise differentiation of degree of illness.

Cutoff scores were also determined in the same manner from the data of this cross-validation sample using the medians reported in Table 2. Prediction levels were enhanced by this method. For scoring category *PO*, 97 per cent of normal *Ss* had 0, 90 per cent of neurotic *Ss* had 1, and 93 per cent of psychotic *Ss* had 2 scores. For *PR*, 100 per cent of normal *Ss* had 0, 96 per cent of neurotic *Ss* had 1, and 93 per cent of psychotic *Ss* had 2 scores. *PP* prediction scores were not appreciably different from those obtained using cutoff scores derived from the original study.

It is suggested that cross-validation differences may result from the use of an outpatient neurotic group which exaggerated the differences between groups by providing *Ss* more nearly equidistant on a hypothetical

Table 3

Percentage of *Ss* in Each Group Receiving Scores of 0, 1, 2 on Each Category

Group	Category *PO*	*PR*	*PP*
Normal			
0	80	98	72
1	20	2	28
2	0	0	0
Neurotic			
0	1	2	10
1	66	81	64
2	33	17	26
Psychotic			
0	0	0	6
1	6	16	57
2	94	84	37

continuum of "illness" between the normal and psychotic groups. This suggestion is given credence by the almost complete absence of overlap between the normal and psychotic groups in the original study. To a considerably lesser degree, the additional training and experience of scorers reflected in the higher reliability figures may have been operative in sharpening differences between groups. The fact remains that there was minimal overlap between the three groups for *PO* and *PR* scores.

Discussion

The use of objective TAT scores for male groups seems to have considerable diagnostic power. This success may be attributable, at least in part, to test scoring based on a *rationale* for simple, consistent, theoretically derived objective scores for all projective tests (3). This rationale is based on three aspects of test behavior deemed sufficient for development of objective scoring systems: (*a*) approach to the situation (reflected in the manner standard test directions are followed); (*b*) normality of response (abstractions of structural and content material included by specified percentages of "normal" *S*s); (*c*) rarity of response (those infrequent responses in a "normal" population which appear with significantly greater frequency in psychopathological conditions). The direct relationship between these three aspects of test behavior and the TAT scoring categories *PO*, *PR* and *PP* is clear. Certain controversial theoretical assumptions concerning personality, i.e., Personality Orientation (1), antedate the rationale from which these scores were developed. However, the use of these categories demands no particular theoretical approach to personality. Similarly, the development of objective TAT scores does not imply abandonment of content analysis but is merely a formal aid to this process.

However, objective TAT scoring to attain more than mere passive clinical recognition must yield descriptive personality data. Further research must concentrate upon this aspect of validity.

Received June 1, 1955.

References

1. Dana, R. H. Personality orientation: an organizational focus for current research. *J. Psychol.*, 1954, **37**, 139–150.
2. Dana, R. H. Clinical diagnosis and objective TAT scoring. *J. abnorm. soc. Psychol.*, 1955, **50**, 19–25.
3. Dana, R. H. The objectification of projective techniques: rationale. *Psychol. Rep.*, 1955, **1**, 93–102.
4. Dana, R. H. Rorschach scorer reliability. *J. clin. Psychol.*, 1955, **11**, 401–403.
5. Dana, R. H. Selection of abbreviated TAT sets. *J. clin. Psychol.*, in press.
6. Dana, R. H. An application of objective TAT scoring. *J. proj. Tech.*, in press.
7. Fleming, Edith. A descriptive analysis of responses in the Thematic Apperception Test. Unpublished master's thesis, Univer. of Pittsburgh, 1946.
8. Spiegelman, M., Terwilliger, C., & Fearing, F. The reliability of agreement in content analysis. *J. soc. Psychol.*, 1953, **37**, 175–187.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Role Taking in Schizophrenia

Isidore Helfand[1]
Teachers College, Columbia University[2]

Experimental investigations of role taking or the empathic process have been primarily conducted with "normal" *S*s, especially college students. The results of these investigations, notably the studies of Dymond (3, 4, 5) and McClelland (8), suggest that the greater the degree of emotional instability, the poorer the ability of the *S*s to take the role of another. These findings are consistent with the hypotheses offered by Cameron (1, 2) who suggested that a disturbance in role-taking skills is crucial to the development of schizophrenia.

The present experiment was designed to explore role-taking characteristics in schizophrenia. In addition, it was hoped that some light would be cast upon the characteristics of empathic ability, not only in schizophrenia, but in the normal individual. Specifically, two major hypotheses were tested:

1. Schizophrenics, when compared with nonpsychotic individuals, show poorer role-taking skills.

2. Nonpsychotic individuals, when compared to schizophrenics, are more homogeneous in their agreement with one another as to the characteristics of the individual whose roles they are instructed to take.

Procedure

Studies in role taking have used what might be termed the method of personal acquaintanceship. Students sharing a common course in college are asked to rate one another. In a study which seeks to compare schizophrenics and normal individuals, such a procedure becomes impractical. It is especially so when it is desired that the degree of familiarity as well as the type of relationship be controlled.

In the present study, therefore, an autobiography obtained from a former hospital patient was used as a common stimulus. The patient had been hospitalized briefly for one of a series of minor recurring depressions. He was frank, honest, and, in the opinion of the hospital staff, someone with whom one could readily empathize. He was asked to discuss his life with reference to parents, siblings, courtship, marriage, sex, education, and vocation. From this material, an 80 item Q sort was constructed. Items were selected which minimized reactions on the basis of fact; rather, they emphasized the need for inference on the part of the *S*. This procedure also reduced the likelihood of the sort's being a mere reading task.

The Q sort was then administered to the author of the autobiography with instructions to distribute the items so as to reflect most accurately his feelings and attitudes. The result is referred to hereafter as the Criterion sort.

Each of the *S*s was asked to take the Q sort twice: First, they sorted the items to reflect their own attitudes and feelings. The second time, they were read the autobiography. They also had a copy which they could read and refer to. They were then asked to take the Q sort again, but this time as it reflected the ideas of the author of the autobiography. They were to act like him when they took the test. The present paper is con-

---

[1] The present paper is a condensation of a part of a doctoral dissertation completed at Teachers College, Columbia University. The author wishes to express his indebtedness to the chairman of his dissertation committee, Professor E. J. Shoben, Jr., and to Professors L. F. Shaffer and Herbert Solomon for their guidance and assistance.

[2] Now at Park Lane Neuropsychiatric Clinic, Cleveland.

Table 1

Summary of Background Data

	C's	P's	T.B.	Normal	F Ratio†
Age in years	35.9	38.2	37.9	33.4	.99
School grade completed	10.6	11.5	10.1	12.2	3.61*
Reading grade level	8+	8+	8+	8+	
Reading score	14.9	17.4	18.5	19.6	17.14**
Vocabulary raw score	23.2	26.5	23.6	25.1	1.3
Hospitalization months	43.9	37.2	37.8		1.06

† The F ratios are the results of the analyses of variance on each of the five factors indicated.
* Significant at the .05 level of confidence ($F_{.05} = 2.76$).
** Significant at the .01 level of confidence ($F_{.01} = 4.13$).

cerned with this second or Simulated sort and its relationship to the Criterion.

Subjects

All 64 *S*s were given screening tests consisting of the Wechsler-Bellevue Vocabulary Test (12) as an estimate of verbal abilities and a reading test (10). The aim here was to eliminate illiterates, functional mental defectives, and those too uncooperative or inattentive to manage the tasks adequately.

Of the 200 male schizophrenic patients at a chronic treatment center between the ages of 20–45 who had been hospitalized for two to five years, 25 remained after screening. These 25 patients fell into two groups, a chronic group of 15 patients, and a privileged group of 10 patients. Chronic patients are confined to the wards and have a poor prognosis. Their activities are limited to simple routine ward chores. Privileged patients have made a sufficient recovery to be permitted freedom of the grounds. They worked in the hospital shops and some were being considered for discharge.

A group of 19 tuberculous patients were included as a control for any possible deterioration due to hospitalization. They were nonpsychotic and had never required consultation for a mental disorder. They had been hospitalized for a period ranging from twenty months to five years.

The normal group included 20 individuals who were functioning members of their community. None had been hospitalized for a mental disorder or had ever sought professional help for such a disturbance. They were firemen, machine operators, clerks and the like. All were gainfully employed.

Table 1 provides a summary of the background data of the subjects employed. An analysis of variance indicated that the groups differed significantly on two factors: Education (at the .05 level of confidence) and reading ability (at the .01 level of confidence). Chronic schizophrenics and tuberculous patients appear to be less well educated than the other two groups. While all groups were able to surpass requirements for an eighth grade reading level, chronic schizophrenics show a marked impairment in these skills, possibly as a function of a concentrative difficulty. Age, length of hospitalization, and verbal abilities do not differ significantly among the groups.

The influence of education and reading ability was considered in the analysis of the results.

Results

Empathic ability was determined by comparing the Simulated sort with the Criterion sort. The individual correlations were converted to z scores since correlations are not normally distributed. The resulting array was subjected to an analysis of variance. The F ratio obtained ($F = 11.05$) indicates that the between group variance was greater than the within group variance to a degree in excess of that which might be anticipated by chance ($F_{01} = 4.13$). The analysis was continued using a t test, the results of which can be seen in Table 2.

As can be seen, all groups differ significantly from the chronic schizophrenic group in their ability to empathize. The privileged schizophrenic group is superior to all others.

Table 2

Mean Empathy Scores of All Groups and the Significance of the Differences Between Means

	C's	P's	T.B.	Normal
Mean	.050	.437	.356	.290
SD	.236	.165	.149	.153
C's *t*		4.61**	4.25**	3.33**
P's *t*			1.2	2.22*
Normal *t*			1.3	

* Significance at or better than the .05 level of confidence.
** Significance at or better than the .01 level of confidence.

This superiority is particularly striking when the privileged patients are compared with normal individuals. Tuberculous patients, while significantly different from the chronic schizophrenics, stand between the privileged and normal groups and are significantly different from neither.

To determine whether these results were influenced either by reading ability or education, an analysis of these two factors was carried out using the Mann-Whitney *U* test. The entire group was split in two, a high and low empathic group. A comparison was then made between poorer readers and more proficient readers. A similar analysis was made comparing education. A critical ratio of .49 for reading ability and .36 for education suggested that these factors had little bearing on empathic ability.

The second hypothesis was tested by correlating the Simulated sort of each individual with the sort of every other individual in his group. The distribution of correlations and their medians can be seen in Table 3.

From this array, it would appear, inspectionally, that the differences between the chronic schizophrenics and other three categories are particularly pronounced. Normals, it would appear, are superior to all groups, and, as was the case with role-taking measures, privileged schizophrenics most closely resemble tuberculous patients. The tuberculous appear to have greater homogeneity than the psychotics. Consequently, there seems to be an increasing degree of homogeneity in

Table 3

Distribution of Intragroup Scores

	C's		P's		T.B.		Normal	
Scores	*N*	%	*N*	%	*N*	%	*N*	%
−.30	1	1						
−.25	3	4						
−.20	8	11						
−.15	6	17	1	2				
−.10	4	20						
−.05	10	30			1	.6	1	.5
0	*34*	*63**			2	2		
.05	19	80	3	9	3	4	5	3
.10	8	88	3	16	9	8	6	6
.15	5	93	8	33	17	19	10	12
.20	4	97	4	42	17	29	15	19
.25	2		*7*	*58**	29	45	14	27
.30	1		3	64	*13*	*53**	9	31
.35			2	69	10	59	18	41
.40			7	85	16	68	*30*	*57**
.45			3	91	14	77	38	77
.50			1	93	15	86	19	87
.55			1	96	11	92	8	91
.60			2		6	95	12	97
.65					2	96	3	99
.70					4	99	1	
.75					1		1	
.80					1			

* Median rank for each group has been italicized.

going from the chronic schizophrenics through the privileged and tuberculous *S*s to the normals.

The fact that each individual was correlated with everyone else within his group makes for a lack of independence in observation that precludes the application of a statistical analysis or any conclusive tests of homogeneity among the groups. While, therefore, it appears that normal individuals are more homogeneous in their agreements and that schizophrenics agree less among themselves concerning the individual whose roles they take, this conclusion must be taken only as suggestive in view of the inapplicability of rigorous tests of significance.

Discussion

The findings tend to confirm certain aspects of the hypotheses. Others remain moot and unsupported. The normal individual's reaction appears to be more consistent with the hypothesis offered by Lindgren and Robinson (7). Rather than responding with any marked degree of sensitivity, normals appear to respond more in terms of a preconceived and more universally shared idea of what another should be like. It seems likely that "well-adjusted" individuals of Dymond's studies had a good appreciation of cultural norms and expectancies from which their judgments were derived. In any event, this factor of conventionality appears to be more important in the normal individual's reaction to another than does the other's idiosyncrasies.

The chronic schizophrenic is apparently deficient in both areas. He is indeed the individual, responding on the basis of his own fantasies and ideas, with little concern for either the accuracy of what he perceives or whether his ideas are shared with others. One limitation of this study is the lack of certainty as to whether the chronic schizophrenic patient responded because of a basic impairment of role-taking skills or because of disinterest in the task. It was hoped that this latter was minimized by screening. Following an analysis of the results, the question appeared to remain unanswered. Because of practical considerations due to the lapse of time, the reliability of the sorts could not be obtained.

Some suggestion that role taking may no be deficient in the chronic schizophrenic, bu merely not utilized, is offered by the result of the privileged patients. Assuming that th latter were once as severely disturbed (requiring custodial treatment) as were th chronic schizophrenics, partial remission fron the disease appears to be characterized by hypersensitivity to the feelings of others, phenomenon Fromm-Reichman aptly title "emotional eavesdropping" (6). In contras to normal individuals, however, they appea to lack a conventional frame of reference Despite their relative accuracy in role taking, each responded to the cues given in a very individual manner.

The results pose at least two problems: *What* is it that permits, or enables the schizophrenic to respond with such sensitivity? Secondly, *why* do they respond with such sensitivity?

In reviewing the method, the task appears to be not unlike a projective or semistructured situation. An opportunity was provided for interpretation rather than a repetition of facts, as so much of the behavior of the *S*s might be characterized as projective in the psychometric sense of the term. The behavior of the normal individual was to project a preconceived set of ideas based primarily on a commonly shared cultural stereotype. The privileged patients, responding to the same information, made better use of it, although their reactions were highly idiosyncratic. Sarbin's description of the schizophrenic as lacking a "generalized other" concept seems appropriate to the results here (11). The schizophrenic, possibly because of this lack, responded to the cues as he perceived them.

The normal individual, on the other hand, may have received the same cues, tested or evaluated them, and, in the absence of corroborating data, rejected many of the hypotheses. The schizophrenic appears to be much less critical. In all probabilities, both this lack of criticalness and a lack of a conventional frame of reference, or what Mead has termed a "generalized other" (9), serve to contribute to the hyperacuity noted.

Assuming that schizophrenics lack this generalized other, such an impairment, with resultant hyperacuity, should be found in younger children, and in emotionally disturbed children as compared to well-adjusted children of the same age. Since such concepts as a generalized other are predicated on the transition from concern with particular people, via interpersonal relationships, disturbances in such relationships, as is indicated in schizophrenia, would contribute to a failure to make the transition. If inconsistency, rejection, and hostility characterize the history of the individual with a predisposition to the development of schizophrenia, he would look on each individual not unlike a sentry on a hostile frontier, carefully evaluating and appraising each one, and never, as do normal individuals, taking others for granted, and relying on generalized norms of what others are like.

Concerning the tuberculous group, the findings offer only equivocal suggestions. In terms of the results, these *S*s appear to most closely resemble the privileged schizophrenic. Whether this is a function of the debilitating social consequences of long hospitalization and the nature of the routine, or of some "tuberculous personality," awaits further research. It is interesting to note that one would expect a higher degree of homogeneity than appeared to characterize these individuals, since they are sequestered together for a long period of time, and have an excellent opportunity to know one another.

Summary and Conclusions

Schizophrenics, tuberculous, nonpsychotic patients, and normal individuals were tested on their ability to take roles. Empathy was determined by their ability to simulate the test performance of an author of an autobiography which was provided them. In addition, the relative agreement with one another, within each group, was determined. The results suggest the following:

1. Chronic schizophrenics have impaired role-taking skills and are relatively individual in their perception.

2. Normal individuals, while superior to chronic schizophrenics in role-taking skills, are more inclined to rely on a conventional frame of reference, rather than demonstrating marked role-taking skills.

3. Privileged patients show a marked sensitivity to others, but appear to respond in a highly idiosyncratic manner.

4. A suggestion was offered that schizophrenics appear to lack a concept of a "generalized other," and that they never make the transition in role-taking skills, from empathy with particular people to a more conventional—universally shared—frame of reference.

5. Tuberculous patients appear to be more like the privileged or partially remitted schizophrenics. The findings here, however, can only be suggestive in the absence of statistical significance.

6. The autobiography appears to be a useful tool in the measurement of empathic ability.

Received June 6, 1955.

References

1. Cameron, N. The paranoid pseudo-community. *Amer. J. Sociol.*, 1943, **49**, 32–38.
2. Cameron, N. *The psychology of behavior disorders*. Boston: Houghton Mifflin, 1947.
3. Dymond, Rosalind F. A preliminary investigation of the relations of insight and empathy. *J. consult. Psychol.*, 1948, **12**, 228–233.
4. Dymond, Rosalind F. A scale for the measurement of empathic ability. *J. consult. Psychol.*, 1949, **13**, 127–133.
5. Dymond, Rosalind F. Personality and empathy. *J. consult. Psychol.*, 1950, **14**, 343–350.
6. Fromm-Reichman, Frieda. *Principles of intensive psychotherapy*. Chicago: Univer. of Chicago Press, 1950.
7. Lindgren, H. C., & Robinson, Jacqueline. An evaluation of Dymond's test of insight and empathy. *J. consult. Psychol.*, 1953, **17**, 172–176.
8. McClelland, W. A. A preliminary test of role-playing ability. *J. consult. Psychol.*, 1951, **15**, 102–108.
9. Mead, G. H. *Mind, self and society*. Chicago: Univer. of Chicago Press, 1934.
10. Monroe, W. S. *Monroe standardized silent reading test, revised, test II, grades 6–8*. Bloomington, Ill.: Public School Publishing, 1920.
11. Sarbin, T. R. The concept of role taking. *Sociometry*, 1943, **6**, 273–285.
12. Wechsler, D. *The measurement of adult intelligence*. Baltimore: Williams & Wilkins, 1944.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Some Psychological Correlates of Humor Preferences[1]

Rudolf Grziwok
The University of California

and Alvin Scodel
The Ohio State University

One hundred and forty male college students were given the following sequence of procedures: (*a*) A series of 40 cartoons on which ratings of funniness were made; (*b*) The TAT (Cards 4, 6BM, 7BM, 13MF, 14, 18BM, 16); (*c*) The Allport-Vernon-Lindzey Study of Values.

The 40 cartoons were selected from a larger group of 250, all of which had originally appeared in *The New Yorker* magazine. Three judges placed each cartoon into one of five categories defined as follows: 1. Humorous effect based on aggression, either explicit or deliberately understated; 2. Humorous effect obtained by a parody on sex; 3. Humor based on the exaggeration or paradoxical use of social stereotypes; 4. Humorous effect based on obvious and striking logical incongruity; 5. No category was applicable or two or more categories were equally applicable.

The first two categories (aggression and sex) can be subsumed under "orectic humor" and three and four (social commentary humor and humor based on logical incongruity) under "cognitive humor."

Seventy-three cartoons, excluding agreements on category 5, were placed in the same category by all three judges, a degree of agreement exceeding chance at the .001 level. The final series of 40 cartoons contained 10 cartoons from each of the first four categories. No cartoonist was represented in any category more than twice, and duplication of themes was avoided as much as possible.

The TAT stories were scored according to specially constructed scales for degrees of aggressive and sexual content and, in addition, intraception vs. extraception. Independent scoring by the writers of a portion of the stories resulted in agreements of 90%, 92%, and 88% respectively; in view of this high agreement, remaining scores were scored by the senior writer.

In the evaluation of results the ratings of funniness were converted to standard scores and then ranked. The TAT and Allport-Vernon-Lindzey scores were dichotomized and then compared on the basis of humor scores by the use of the Mann-Whitney test.

The significant results indicated that subjects high in TAT aggression prefer aggressive humor while those low in TAT aggression prefer social commentary humor. With respect to value orientations subjects high on the aesthetic scale prefer logically incongruous cartoons whereas those who are low on the same scale prefer aggressive humor. Moreover, subjects high on the social scale prefer aggressive cartoons and those low in the theoretical value show a preference for sexual cartoons. In more general terms a preference for orectic humor, as opposed to cognitive humor, seems to be characterized by more fantasy aggression, more extraversion or outgoingness, less preoccupation with intellectual values, and less psychological subtlety or complexity.

Brief Report.
Received October 10, 1955.

---

[1] An extended report of this study may be obtained without charge from Alvin Scodel, Psychology Department, The Ohio State University, Columbus 10, Ohio, or for a fee from the American Documentation Institute. Order Document No. 4742 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.75 for microfilm or $2.50 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

A Dimensional Analysis of Depth of Interpretation[1]

H. L. Raush, Z. Sperber, D. Rigler, J. Williams, N. I. Harway,[2]
E. S. Bordin, A. T. Dittmann,[3] and W. L. Hays
University of Michigan

The research worker interested in problems of human interaction is faced with a dilemma. The concepts he is concerned with are complex ones. If he chooses in his research to simplify the variables he employs, so as to minimize problems involving issues of measurement, the assumptions he must make obscure the relationship of his variables to the concepts and theories he started from. If, on the other hand, he prefers to work with variables directly relevant to his theoretical formulations, he must face a multitude of issues in the course of developing adequate methodological tools (2).

This dilemma has come to our attention particularly in our studies of depth of interpretation. This complex variable derives directly from psychoanalytic theory (6) and has been a subject of disagreement among therapeutic theories (3, 5). As a necessary step toward testing the differing theoretical viewpoints about depth of interpretation, we have developed a measure for the concept, using conventional methods for constructing a rating scale (8). However, a rating scale such as we have used and such as is common in other clinical studies, forces judges to respond as though they were dealing with a single dimension, irrespective of the true complexity of the variable. Therefore, the dimensionality of depth of interpretation remains to be studied.

We shall describe four studies designed to determine: (*a*) whether depth of interpretation is uni- or multidimensional; (*b*) whether the dimension or dimensions revealed are consistent with our concept of depth of interpretation; (*c*) to what extent other dimensions can be identified.

If we demonstrate that depth of interpretation is in fact unidimensional, we can proceed to use the scale and to make more confident inferences about its relationship to other variables. However, interjudge reliabilities obtained for rating of single therapist responses were low enough to raise questions about the likelihood of finding unidimensionality. If we uncover evidence of more than one dimension, but we judge that these dimensions can be incorporated into a revised notion about depth of interpretation, then our study will have brought about a modification in our concept or theory. Also, by being able to identify these dimensions we may be able to achieve better agreement among raters. Finally, if more than one dimension is revealed, any of the additional ones may prove irrelevant from our point of view and might be susceptible to elimination through refinement of the scale or through training of judges.

In summary, our specific purpose in this paper is to find out what attributes judges use in dealing with depth of interpretation, to what extent they agree as to the attributes, and to what extent any of these attributes corresponds to our definition and scale of depth of interpretation. The approach we shall discuss is applicable to a wide variety of clinical studies. It accepts the complexity of psychological phenomena; it places relatively little constraint on the data; and it allows us to estimate the number and to describe some characteristics of dimensions that judges are actually using in evaluating complex attributes.

---

[1] This paper is part of an investigation supported by a research grant M-516 C-2 from the National Institute of Mental Health of the National Institutes of Health, United States Public Health Service. Principal Investigators are: E. S. Bordin, H. L. Raush; Coordinator for the Project: Z. Sperber. Part of this paper was presented before the Midwestern Psychological Association, Columbus, Ohio, May, 1954.

[2] Now at the University of Rochester.

[3] Now at the National Institute of Mental Health, Bethesda, Maryland.

General Method

Definition. Depth of interpretation was defined for the judges as follows: "any behavior on the part of the therapist that is an expression of his view of the patient's emotions and motivations—either wholly or in part—is considered an interpretation. A patient has varying degrees of awareness of his emotions and motivations. Depth of interpretation is a description of the relationship between the view expressed by the therapist and the patient's awareness. The greater the disparity between the view expressed by the therapist and the patient's own awareness of these emotions and motivations, the deeper the interpretation." In most of the studies the judges were given, in addition to the definition, a seven point scale with nine descriptive statements illustrative of various points on the scale (8).

Judges. The judges used for all studies were Ph.D.'s, staff psychologists, or advanced graduate students at the University of Michigan, all of whom had at least 100 hours of experience as psychotherapists.

Stimuli. All stimuli consisted of excerpts from transcripts of recorded therapeutic sessions. Judges were presented with three consecutive patient-therapist exchanges, ending with a therapist's remark. Only the final therapist response was to be judged; the preceding section served to provide context. Selection of the stimuli will be discussed with reference to each study.

Technique of data collection and analysis. Since the methods used in this study are of recent origin, it seems appropriate to provide a brief summary of them. Bennett (1) has developed techniques which enable us to estimate the number of dimensions in data that are collected by methods developed by Coombs (4). Hays (7) has expanded the procedure so as to provide information on the ordering of the stimuli on the several dimensions. These methods enable us to investigate the underlying factors entering into judgments of behavior by providing a set of dimensions which best account for a given set of judgments.

All possible combinations of the stimuli, taken three at a time, were presented. The task for the judge was to say which two of the three stimuli were most alike and which two were least alike with respect to depth of interpretation (Method of Similarities). In the first two studies where six stimuli were used, there were 20 such triads; in the third and fourth studies using seven stimuli, there were 35 triads. These triads were presented in a randomized order. Using each stimulus of the triad in turn as the reference point, we can then rank the other two with respect to their psychological distance from this referent.[4]

All possible paired comparisons are derived in association with each reference point. If the judgments are transitive—that is, if the judge behaves consistently in going from triad to triad—the partial rank orders derived from the separate triads can be combined to form a complete rank order, or I Scale, with the reference stimulus coming first in the rank. Thus for each judge, the original judgments of similarity potentially can yield one I Scale from the "point of view" of each stimulus. The total number of I Scales that can be obtained from any one judge is therefore equal to the number of stimuli used in the study. By means of Coombs' Unfolding Technique (4), we can then determine whether the I Scales for each judge are compatible, that is, whether they could all have arisen from a single dimension.

Even if judges differ in their I Scales, if we assume that in making their comparisons they operate from some segment of a larger space in which the stimuli are located, we may pool all of the I Scales. Bennett's techniques (1) then allow us to estimate the number of dimensions required to account for the observed I Scales, and Hays' techniques (7) may be applied to locate and order the stimuli in this larger space.[5]

---

[4] For example, if in judging the three Stimuli A, B, and D, the judge decides that A and D are most alike and B and D least alike with respect to depth of interpretation, we would convert these decisions into rankings as follows: From the "point of view" of stimulus A the ranking would be ADB; from the "point of view" of stimulus B the ranking would be BAD; from the "point of view" of stimulus D the ranking would be DAB.

Study 1

The first study was divided into two parts. In the first part we tried to determine whether 10 judges could judge depth of interpretation unidimensionally when it was defined for them with only the paragraph (quoted above) regarding interpretation and depth. In the second part we wanted to determine whether and how the judgments and dimensionality would be modified when the same 10 judges were asked to repeat the task, but this time were also given as a guide the graphic rating scale (8).

Selection of stimuli. Six stimuli were selected, three from each interview, from two of a series of interviews involving different therapists and patients. Therapist responses in these interviews had been rated for depth of interpretation as part of another study (8). The stimuli were selected so that mean ratings varied across the range of depth of interpretation, and so that the variance of the ratings for five out of the six therapist responses was at or below the average item variance for the interview.

Results. When judges had only the definition available to them, 35 of the 60 possible I Scales were obtained and no judge had a complete set of compatible I Scales. Since a complete I Scale can be recovered from the data only when the judge's comparisons are transitive, and since the comparisons will be transitive and his I Scales compatible only when he treats the concept in a unidimensional manner, it is evident that no judge handled the concept of depth of interpretation unidimensionally. At least three dimensions were necessary to account for the 35 obtained I Scales, but because of the limited amount of data and the evident confusion of the judges, we could not order the stimuli on these dimensions.

Where both the definition and the graphic rating scale were available to the judges, there were 53 complete I Scales out of the possible 60. The data of three judges now yielded complete and compatible sets of I Scales. In analyzing the pooled data, three dimensions again were required, indicating that despite the addition of pretested points on the scale to the careful definition of depth of interpretation, the judges did not deal with the excerpts as though depth of interpretation were a single continuum.

The stimuli on the first dimension could be clearly ordered and inspection of them allowed us confidently to name the dimension depth of interpretation. The six excerpts could be ordered on the basis of the means of the ratings that had been obtained in the study referred to previously. The rank orders arrived at in these two diverse fashions correlated .89.

Study 2

Was it the improved frame of reference supplied by the graphic scale, or prior practice on the task which led to the improved results on Part 2 of the study reported above? To investigate this, a second study was initiated using the same six excerpts and 10 new judges. These judges did the task only once using both the definition of depth of interpretation and the scale as a frame of reference. In this study 45 out of 60 possible I Scales were recovered, suggesting that the scale items did improve the frame of reference, but that practice probably helped too. Three dimensions were required. The stimuli could be ordered on the first and second dimensions, but only the polar stimuli of the third dimension could be ascertained. The first dimension was clearly depth of interpretation.

To supply more material, the 53 I Scales of Part 2, Study 1, and the 45 I Scales of Study 2 were pooled. Again depth of interpretation emerged sharply as the first dimension, the rank-order correlation with the ordering based on the average of the previous rating data being .94. The third dimension still could not be identified, but the ordering of stimuli on the second dimension suggested either of two interpretations. Judges could be reacting according to the degree of emotionality they anticipated the therapist's response would arouse in the patient (hereafter referred to as emotionality); or judges could be reacting to the degree of ambiguity of the intent of the therapist's response (hereafter referred to as ambiguity).

---

[5] An outline of the method and a step-by-step illustration of the procedure as applied to our data have been prepared by Dr. Hays and will be made available on request.

Study 3

If the secondary factors could be specified more accurately and if they could be demonstrated to be general rather than specific to the six stimuli we happened to use, then judges could be trained to recognize and correct for the interference of such variables with judgments of depth of interpretation. Accordingly, a third study was run.

Here judgments were again made with respect to depth of interpretation. However, since we were interested primarily in clarifying the meaning of the second dimension, stimuli were selected so as to maximize its possible influence. An important proof of the validity of our interpretation of the second factor would be to find that this second factor now becomes the primary source of the ordering of stimuli.

Selection of stimuli. Three stimuli from those previously studied were included. They were the first, middle, and last stimuli on the second dimension. If the same ordering of these "marker" stimuli was not maintained in the present study, we would be forced to reject the hypothesis that a general secondary dimension operated.

In addition to the three "markers," four new stimuli were chosen in the following manner: from three new interviews we selected ten stimuli, which on the basis of inspection seemed to represent various degrees of both emotionality and ambiguity. These ten stimuli and the three "marker" stimuli were ranked for emotionality and ambiguity by six project staff members. The four stimuli were chosen on the basis of the mean rankings, so that ambiguity and emotionality would be independently distributed, and so that a range would be represented on both these variables. The project staff also ranked the final seven stimuli for depth of interpretation. All possible combinations of the seven stimuli, taken three at a time, provided 35 triads for judgment. For each judge there were potentially seven I Scales.

Judges. Only five new judges were available. In order to provide sufficient judgments about the relations among stimuli, the entire series of judgments was made twice. Between each set of judgments on these seven stimuli, the judges made similar judgments on another group of stimuli discussed in Study 4 below. By this procedure we attempted to maximize independence between the two sets of judgments.

Results. Forty-five out of the 70 possible I Scales were obtained. Three dimensions were required to account for these data. The rankings by the project staff provide criteria for determining whether any of these three dimensions could be interpreted as representing either emotionality, ambiguity, or depth of interpretation. Although we expected depth of interpretation to show up less strongly in stimuli chosen for the purpose of carrying the secondary qualities, again the first dimension was depth of interpretation, correlating (*rho*) .93 with the pooled rankings. The ordering of the stimuli on the second and third dimensions was not consistent with the criteria for either emotionality or ambiguity. Inspection of the orderings gave no clue for interpretation of these secondary dimensions. We must conclude that the secondary dimensions operating in judgments of depth of interpretation were specific to the judges and the stimuli studied.

Study 4

To test the generality of the depth of interpretation dimensions, the first, middle, and last stimuli from the first dimension (depth of interpretation) of Studies 1 and 2 were used as "markers," in conjunction with four new stimuli. These new stimuli were selected in the following manner: A sample of seven was taken from the three additional interviews used in Study 3; the stimuli were chosen initially because they seemed to represent the depth of interpretation dimension. These seven stimuli plus the three "markers" were ranked for depth of interpretation by project staff members. The four new stimuli were then selected on the basis of their mean rankings varying across the range of depth of interpretation. The seven stimuli, three "markers" plus four new stimuli, provide 35 triads for judgment and seven potential I Scales for each judge.

Judges. The five judges of Study 3 went through the series of triads twice. As noted above, to maximize the independence of these repeated judgments, work on the stimuli of Study 3 intervened between the two series.

Results. Forty-five out of the 70 possible I Scales were obtained. Two dimensions were required to account for these data. Except for one of the stimuli, a single dimension would have been sufficient. The first dimension was clearly depth of interpretation and agreed (*rho*) .89 with the order of the pooled ranking data. Agreement with the order of the original "markers" was perfect.

Discussion

The researches reported have demonstrated the applicability of a method of data collection and analysis to the study of a variable in psychotherapy—depth of interpretation.

We had been concerned with the possibility that the concept of depth of interpretation, despite the importance attributed to it in theoretical thinking, might, in fact, have only limited utility. However in each of the four studies, involving a sample of stimuli from five interviews with different patients and therapists, depth of interpretation was consistently the primary dimension. Although the method of analysis is not sufficiently developed to allow for the quantification of variance, the logic underlying it permits us to state with confidence that more variability in our data can be attributed to the depth of interpretation dimension than to any of the other dimensions operating.

We had also been concerned with the low interjudge reliabilities observed when we attempted to measure the depth of interpretation of each therapist response with a carefully developed rating scale (8). One possible source of unreliability may be that the use of the rating scale forced judges to treat as unidimensional a complex aspect of interview material involving several dimensions. However, despite the extremely complex nature of the material rated, there does seem to be a significant area of agreement between judges. The high correlations between the recovered depth of interpretation dimensions with pooled ratings or rankings by judges, indicate that ratings may be employed despite low individual reliabilities. That is, in using ratings of depth of interpretation, it may be possible to overcome problems of low reliability of single judgments on this attribute by utilizing average ratings of several judges. Since depth of interpretation, the dimension in which we are interested, contributes most strongly to the judgments, pooling judgments may maximize its contribution and reduce the effect of secondary factors, specific to the ratings of individual judges.

The experimental method outlines a procedure which can be applied to the study of many of the complex variables discussed in psychotherapeutic and clinical research. The assumption of a single dimension, implicit in the construction of a rating scale, often masks the fact that additional factors are influencing a phenomenon. Lack of awareness of such factors often leads to experimental results which cannot be clearly interpreted. The type of research we have reported allows the experimenter to ascertain whether his variable is, in fact, a unidimensional attribute, and, if not, under what conditions the powerful and convenient rating scale technique may legitimately be applied.

Summary

Four studies of the dimensional characteristics of a psychotherapeutic variable, depth of interpretation, were reported. Depth of interpretation was shown to be treated by judges not as a unitary dimension but as at least three dimensions: The primary dimension consistently was depth of interpretation; the secondary dimensions varied depending on the judges and stimuli used. The implications of these results for research with complex psychological variables were discussed.

Received May 9, 1955.

References

1. Bennett, J. F. A method for determining the dimensionality of a set of rank orders. Unpublished doctor's dissertation, Univer. of Michigan, 1951.

2. Bordin, E. S., Cutler, R. L., Dittmann, A. T., Harway, N. I., Raush, H. L., & Rigler, D. Measurement problems in process research on psychotherapy. *J. consult. Psychol.,* 1954, **18,** 79–82.

3. Collier, R. M. A basis for integration rather than fragmentation in psychotherapy. *J. consult. Psychol.,* 1950, **14,** 199–205.

4. Coombs, C. H. Theory and methods of social measurement. In L. Festinger & D. Katz (Eds.). *Research methods in the behavioral sciences.* New York: Dryden Press, 1953. Pp. 471–535.

5. Dittmann, A. T. The interpersonal process in psychotherapy: development of a research method. *J. abnorm. soc. Psychol.,* 1952, **47,** 236–244.

6. Fenichel, O. Problems of psychoanalytic technique. Albany, N. Y.: *Psychoanalytic Quart.,* 1941.

7. Hays, W. L. An extension of the unfolding technique to r-dimensions. Unpublished doctor's dissertation, Univer. of Michigan, 1954.

8. Harway, N. I., Dittmann, A. T., Raush, H. L., Bordin, E. S., & Rigler, D. The measurement of depth of interpretation. *J. consult. Psychol.,* 1955, **19,** 247–253.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Rating Personal Adjustment Through an Analysis of Social Reinforcements

Vaughn J. Crandall and Ursula Bellugi
The Fels Research Institute for The Study of Human Development

Methods for assessing adjustmental aspects of behavior have often been unrelated to any consistent personality theory. As a result, behavior sampled by these methods has frequently been analyzed according to a hodgepodge of empirically derived, and theoretically unrelated, adjustment indicators. Current "sign approach" methods for assessing adjustment reflected in Rorschach responses are a case in point. Such empirically derived indicators of adjustment can seldom be used to analyze behavior evoked by assessment techniques other than those for which they were devised. The aim of the present study was to deal with adjustmental aspects of behavior from the point of view of one theory of personality, and to develop a method of analysis within this context which might be applied to behavior sampled by a variety of personality assessment techniques.

The method of analysis developed in the study derives its theoretical basis from the Social Learning Theory of Personality of Rotter and his students. This theory contains a construct of *freedom of movement* which has been defined as, "The mean expectancy of obtaining positive satisfactions as a result of a set of related behaviors directed toward the accomplishment of a group of functionally related reinforcements . . . " (2, p. 194). The construct of freedom of movement is an abstraction concerning the adjustmental aspects of an individual's behavior in relation to his specific needs or social goals. Inferences concerning an individual's over-all freedom of movement would, of course, necessitate evaluating an individual's behavior in relation to all his major needs or goals. This could be done by ascertaining the relative frequency with which his past behavior directed toward these goals had resulted in success and satisfaction or failure and dissatisfaction. A complete study of an individual's past history of social reinforcements would, of course, be impossible. It is feasible, however, to obtain samples of his current behavior and to arrive at an estimate of his present freedom of movement through the analysis of these behavior samples. The Social Reinforcement Index (SRI) of the present study was developed with this aim in view.

Method

Subjects

The *S*s of the study were 87 mothers of children enrolled in a longitudinal study of maternal behavior and child development at the Fels Research Institute for the Study of Human Development. Most of the *S*s were members of the middle socioeconomic class. Their distribution on the Index of Status Characteristics of Warner *et al.* (3) was: upper-lower class, 17%; lower-middle class, 59%; upper-middle class, 20%; and lower-upper class, 4%. The *S*s were more intelligent and better educated than national averages. Their median Wechsler-Bellevue full-scale IQ was 123. Concerning their educational background, 7 per cent had not completed high school, 29 per cent were high school graduates, and 64 per cent had attended college.

Protocols

While the SRI method of analysis is potentially useful for the assessment of freedom of movement reflected in behavior sampled by a variety of personality assessment techniques, the present study was limited to the investigation of its applicability for the analysis of two separate kinds of protocols: (*a*) a clinical psychologist's descriptions of the characteristic behavior of the *S*s in various areas of their life, and (*b*) interviews with *S*s concerning their behavior in these areas. The former protocols will be called *observer descriptions* and the latter will be designated *subject interviews*.

The observer descriptions were obtained from a clinical psychologist, who, as a Fels Home Visitor, had regularly visited the *S*s in their homes at least twice a year as part of a continuous Fels Home Visit Program investigating maternal behavior and child development (1). In addition, the Home Visitor had observed and interviewed the *S*s from time to time at the Fels Institute. The observer descriptions, based on the Home Visitor's total knowledge of each *S*, covered the following areas: (*a*) general methods of dealing with day-to-day situations, (*b*) housework and home-centered activities, (*c*) maternal behavior, (*d*) marital relationship, (*e*) interactions with friends, and (*f*) social and community activities. These descriptions were spontaneous verbal reports told to the senior author by the Home Visitor without previous preparation. They were electrically recorded and typed for subsequent SRI analysis. They averaged just under five hundred words in length.

The subject interviews were conducted by a second clinical psychologist, also a Fels Home Visitor, who individually interviewed 41 *S*s at the Fels Institute.[1] These interviews covered the same areas of the *S*s' activities as did the observer descriptions. During the interviews standard questions were employed, designed to elicit not only the *S*'s descriptions of experiences in each area, but also her satisfactions and dissatisfactions with these experiences. The interviews averaged one and one-quarter hours in length with the *S*s' verbalizations during the interviews averaging approximately seven thousand words. These interviews were also electrically recorded and typed for subsequent SRI analysis.

Social Reinforcement Index Ratings

While the SRI method of analysis results in a global estimate of an *S*'s over-all freedom of movement, the ratings upon which this estimate is based are not global ones. These were purposefully avoided. Global ratings have the mixed blessing of allowing the rater to base his judgments on the total context of the situation as he views it. He is free to attend selectively to certain behavioral referents and to minimize or ignore others. He weighs the importance of a given instance of behavior as he sees fit. Such ratings are thus dependent upon the rater's personal system of "clinical beta weights." The moderate to poor interrater agreement frequently obtained with global ratings may sometimes have been the result of the individualistic nature of this process.

One potential answer to this problem, and the one used in SRI analysis, is the utilization of specific, relatively small, yet psychologically meaningful, rating units. With such a method, the rater is not free to base his final judgment on his idiosyncratic decisions concerning the relative importance of various aspects of the behavior he is rating. Rather, his sole task is to note when scorable behavior occurs, and to categorize the nature of this behavior. Such a process simplifies the rater's task, yet still allows him to make psychologically meaningful judgments.

The scoring unit of the SRI was defined as, "any reported attitude or behavior of the *S* or of persons with whom he interacts which indicates the effect of this experience on the *S*." The unit was scored positive when it contained an experience which was judged satisfying or rewarding for the *S*; the unit was scored negative when the social reinforcement was judged to be an unsatisfying, nonrewarding or frustrating one. In general, raters were instructed to rate only behavioral referents from which they could directly infer satisfaction or dissatisfaction on the part of the *S* being rated. However, occasionally, when the satisfaction or dissatisfaction of the *S* was not directly inferable, it was possible to judge the expected effect of the experience upon the *S*. On these occasions, the raters judged the potential positive or negative social reinforcement inherent in such behavior in terms of the usual reinforcement value for this kind of behavior within the *S*'s cultural groups.[2]

---

[1] The authors would like to express their appreciation to Fels Home Visitors, Dr. Joan Lasko and Mrs. Anne Preston, for their participation in the study.

The SRI was developed to simplify the rating process and should, if it were successful in its aim, be usable by raters with little psychological training and experience as well as by experienced clinicians. To test this, psychologically inexperienced persons were used as SRI raters in the present study. The observer descriptions and the subject interviews were rated by two senior undergraduate students majoring in psychology.[3] These raters had no contact with, or knowledge of, the *S*s whose protocols they rated. A training session preceded the actual rating of the protocols. During the training session, the SRI method of analysis was explained to the raters by the senior author, the raters made practice ratings on a half-dozen protocols not used in the study, and these practice ratings were discussed in relation to original scoring instructions. The raters then independently rated the protocols of the study. The final SRI score for each protocol was obtained by subtracting the sum of the units scored for negative reinforcement from the sum of the positive reinforcement units, and dividing this by the total number of scored units.

Results

The SRI ratings of the protocols were compared with Home Visitors' ratings of the *S*s' over-all freedom of movement. The first Home Visitor rated the 72 *S*s she had discussed in the observer descriptions, and the second Home Visitor rated the 41 *S*s she had interviewed. Each Home Visitor based her ratings on her total knowledge of each *S* gained from all observations and interviews of that *S* in her home and at the Fels Institute. These ratings were thus based on relatively extensive knowledge and made by experienced clinical psychologists. The SRI ratings, in contrast, were made by undergraduate students with little psychological experience, and were based solely on the *S*s' protocols.

The junior author's SRI scores of the observer description protocols were correlated with the first Home Visitor's ratings of the *S*s she had described.[4] An *r* of .81 was obtained, indicating high agreement between SRI analyses of the descriptions and the Home Visitor's judgments of the *S*s' over-all freedom of movement. The junior author's SRI scores of the 41 *S*s' verbalizations during the subject interviews were correlated with the second Home Visitor's ratings. The *r* was .67; SRI analyses of the *S*s' responses during a single interview bore considerable relationship to the *S*s' observed freedom of movement as rated by the second Home Visitor.

Information concerning the reliability of the SRI ratings was also obtained. Forty of the observer descriptions were randomly picked, and the two SRI raters' scores for these protocols were correlated. The correlation was .96. The correlation between the SRI scoring of the two raters on the 41 subject interview protocols was .87. Intrarater reliability of SRI scoring was obtained by comparing the original ratings of the junior author with reratings done one month later. The correlations were .96 for the observer descriptions and .92 for the subject interviews.

Discussion

The general results of the study are encouraging. The SRI analyses of the protocols resulted in reasonably reliable estimates of the *S*s' over-all freedom of movement. In addition, even with psychologically inexperienced raters, inter- and intrarater agreement was high. These conclusions, of course, must be limited to the kinds of protocols analyzed in the study. While the SRI appears to be a fruitful method for analyzing such diverse data as TAT stories, incomplete sentence responses, autobiographical material, or time samples of ongoing social behavior, its applicability for these data must be assessed in future research. Finally, it should be pointed out that the use of the SRI is not necessarily restricted to the evaluation of over-all freedom of movement as in the present study. It may also prove useful, both clinically and experimentally, for the assessment of freedom of movement in respect to specific needs, value areas or social roles.

---

[2] Mimeograph copies of examples of reinforcement units scored for positive and negative reinforcement in the observer descriptions and subject interviews may be obtained from the authors.

[3] The Social Reinforcement Index raters were the junior author and Miss Barbara Fisher.

[4] All correlations reported are Pearson *r*'s. Distributions of scores upon which these correlations were based were all normal, and all relationships were rectilinear by inspection.

Summary

The aim of the study was to deal with adjustmental aspects of behavior from the point of view of one theory of personality, and to develop a method of analysis within this context which might be applied to behavior sampled by a variety of personality techniques. The Social Reinforcement Index was developed for this purpose.

The Social Reinforcement Index method of analysis was employed by two raters to analyze independently two kinds of protocols: (*a*) a psychologist's descriptions of *S*s' behavior in various areas of their life, and (*b*) recorded interviews with *S*s concerning these areas. The Social Reinforcement Index scores derived solely from the ratings of these protocols by relatively inexperienced raters were compared with ratings of the *S*s' freedom of movement by psychologists who had extensive experience interviewing and observing the *S*s. Correlations were high. Social Reinforcement Index ratings also had high interrater and intrarater reliability.

Received May 13, 1955.

References

1. Baldwin, A. L., Kalhorn, Joan, & Breese, Fay. Patterns of parent behavior. *Psychol. Monogr.*, 1945, 58, No. 3 (Whole No. 268).
2. Rotter, J. B. *Social learning and clinical psychology*. New York: Prentice-Hall, 1954.
3. Warner, W. L., Meeker, Marchia, & Eells, K. *Social class in America*. New York: Stratford Press, 1949.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

A Factorial Study of Personal Constructs[1]

Leon H. Levy
Indiana University

and Robert D. Dugan
Commonwealth Life Insurance Company

The purpose of this paper is to present a method for the delineation of an individual's conceptual structure. It has long been an accepted tenet of projective psychology that each individual imposes his own unique structure on the world about him. Accepting this as a basic assumption, one is next required to tease out the dimensions utilized by the individual in this structuration. This is a common goal of the various extant projective techniques. But whereas in these techniques the final delineation rests rather heavily upon an inferential structure which may vary widely from one test analyst to another, it is proposed that the technique to be described below reduces such inferential variability to a minimum.

The Role Construct Repertory Test and Underlying Assumptions

The following assumptions are believed to be the minimum necessary for the derivation of the present technique:

a. For each individual there exists a universe of persons which constitutes his social environment.

b. Each individual possesses a repertoire of constructs which is relatively stable over a period of time, and which he utilizes in structuring his social environment.

c. Constructs contained in a given individual's repertoire bear a relationship to each other such that they may be ordered to certain basic dimensions which define the parameters of his construct repertoire.

d. The structure of an individual's social environment may be duplicated by an observer through knowledge of the parameters of his construct repertoire.

An individual's constructs are viewed in much the same fashion as the constructs of the scientist. They are the ways in which order is brought to the unordered, and the bases upon which predictions about outcomes of behavior are made. Defining a construct as *a way in which two or more things are alike and at the same time different from one or more other things*, Kelly (2) has developed the Role Construct Repertory Test (RCRT) as a means of eliciting an individual's construct repertoire.

One form of the RCRT consists of two parts: In the first part there is a list of 15 *role titles* such as: Your mother or the person who played the part of your mother; A son that liked you; A teacher for whom you had a great deal of respect; etc. For each of these titles *S* writes down one name and may not use the same name twice. In the second part *S* is presented with 15 combinations of these people taken three at a time and is asked to indicate in a word or phrase how any two in each triad are the same and at the same time different from the third. This constitutes a construct. He is then asked to indicate what he feels to be the opposite of the construct he has just listed. Thus there are elicited 15 constructs and their opposites. Using college students and a slightly different form of the RCRT, Hunt (1) obtained a test-retest agreement of 70 per cent on constructs used, indicating some generality for the constructs elicited.

---

[1] This study draws heavily upon the Personal Construct Theory of George A. Kelly (2), although the authors accept full responsibility for the assumptions and present form of this paper. The data upon which this article is based were collected while the authors were in residence at the Ohio State University.

Having obtained a sample of the individual's construct repertoire the next problem involves ordering these to certain dimensions which may be said to represent the parameters of his construct repertoire. It will be apparent to the reader at this point that this is a similar problem to that found in psychometrics, where one has a large number of scores on various tests and wishes to reduce these to more meaningful dimensions which will explain whatever relationship exists between these scores. The method of choice in psychometrics appears to be factor analysis and this method has been applied to the present study. Thus we propose to present the application of factor analysis to an individual's RCRT protocol as a means of determining the manner in which he structures his social environment.

Although some similarity to Osgood's method of deriving his Semantic Differential (4) will be noted, one important difference is that the dimensions arrived at by the present approach are unique for each individual, whereas the scales contained in the Semantic Differential are nomothetic in nature, having been derived from a large group of *Ss*. While Osgood's procedure permits the specification of the meaning of concepts in a multidimensional space defined by his instrument, we would claim that the factor analysis of an RCRT protocol permits, first, the specification of a multidimensional conceptual space unique for a given individual, and second, the location of persons in this space.

Procedure

The RCRT as described above was administered to each of four *Ss*.[2] The resulting constructs in each protocol were then designated as five-point rating scales on which *S* was to rate each of the 15 individuals named in the first part of the test. The low end of the scale meant that the construct was very typical of the person, while the high end meant that the contrasting or opposite construct was very typical. In order to reduce any halo effect, the *S* was asked to consider one construct at a time, rating all of the individuals in succession.

One can thus conceive of the constructs as tests on which there are as many scores as there are people rated. It is recognized that the correlations which result from this procedure will be affected by the fact that the individuals rated on a particular construct may have also entered into the formulation

Table 1

Constructs and Orthogonal Factor Loadings for Case I

Constructs		Factor loadings				
Construct	Contrasting construct	I	II	III	IV	h^2
1. unassuming	pretentious	−19	90	27	−09	93
2. creative	uncreative; static	24	27	90	12	95
3. rigid	flexible	03	−55	−55	−53	89
4. original; daring	unoriginal; conventional	03	−01	40	58	50
5. sincere; unaffected	insincere; affected	−19	91	33	07	98
6. aggressive	submissive	13	−58	−59	30	79
7. sexual	asexual	84	00	03	10	72
8. intelligent	stupid	06	29	92	02	93
9. refined	crude	00	87	31	19	89
10. intellectual	non-intellectual	15	12	89	25	89
11. non-productive	productive	05	−02	−62	−25	45
12. liberal	conservative	40	−17	62	51	83
13. ascetic	sensual	−75	15	04	−52	86
14. introjected values	experience-derived values	−07	−11	−54	−33	42
15. socially inept	socially skilled	05	−14	48	−30	34

---

[2] In order to conserve space only two of the four cases are presented.

of that construct. It was apparent, however, that the correlations were not uniformly affected since they range from high positive to high negative.

The intercorrelations between constructs for each *S* were factor analyzed using the Thurstone multiple group method. Factors were rotated to orthogonal simple structure. The loadings were then adjusted by the Wherry-iterative procedure (5) so that no residuals in any of the tables exceeded ± .15. Ninety-five per cent of the residuals in each of the tables were between ± .10. Loadings of .30 or greater were considered significant.

Results

Since the purpose of this paper is primarily methodological and heuristic, no data will be presented other than the constructs with their factor loadings and a brief description of the *S*.

Case I. A 31-year-old, single male graduate student, considered a "capable" person by his associates. He comes from one of the "old families" of the South and has described his home as somewhat puritanical. He is much concerned with problems of creativity and personal freedom.

Table 1 presents his construct repertoire and orthogonal factor loadings. We consider these factors to be an operational definition of this *S*'s mode of structuring his social environment. To attempt to give names to these factors would appear to be of dubious value, while inspecting each factor for those constructs with significant loadings will provide us with some insight into the ways in which the *S* attempts to make sense out of his world.[3]

Factor I is clearly concerned with sexuality, with the probable identification of sexual expression with liberality and sexual repression with conservatism. One might suspect here that since he describes his family as puritanical, he would perceive them as conservative and hence not condoning sexual expression. The fact that at 31 he is not married may indicate a closer tie to his family than he would be willing to admit. These of course are in the nature of low order hypotheses which such a protocol, taken together with other information, might generate.

Factor II appears to be an affective or attitudinal dimension by means of which people are seen as either unassuming or pretentious with the further implication that persons who are unassuming, submissive, and refined are more to be trusted and relied upon than those for whom the opposite might be true. One would be led to predict here, among other things, that this person might have difficulty relating to persons in certain occupations or social strata where a premium is placed on aggressiveness. Similarly he may be an unduly submissive individual because of the negative implications of aggressiveness. However, he would probably see this more as being flexible than submissive.

Factor III seems to reflect his concern with intellectual freedom and creativity. Interestingly, we also find that the constructs "aggressive–submissive" and "refined–crude" have significant loadings on this factor. Here it appears that creativity and intelligence also carry certain emotional connotations for him and that, perhaps, he is not able or willing to accept a person purely on the basis of his accomplishments in a given area, but also looks for certain nonintellectual components as well.

Factor IV may be the dimension by which he discriminates between people who are like his family and those who are not. If we were interested in transference relationship problems, we might view this factor as mediating the transference. We might expect, if this were true, that having placed a nonfamily figure on the same end of this dimension as his family, he would then transfer to them many of the attitudes he holds in regard to his family. Such information would be invaluable to a clinician working with this person.

Case II. By way of contrast we present the factor analysis of the RCRT protocol of a 48-year-old female graduate student, married, and mother of two children. At the time of this study she described herself as con-

---

[3] While no attempt is made in this article to establish the validity status of the RCRT or our method of analysis, it might be mentioned parenthetically that a faculty member in the department in which this *S* was a student identified the student immediately upon reading this analysis of his protocol, although he had no knowledge that the student had taken part in the study.

Table 2

Constructs and Orthogonal Loadings for Case II

Constructs		Factor loadings				
Construct	Contrasting construct	I	II	III	IV	h^2
1. open mind, growing	closed mind, "set in ways"	91	25	00	30	98
2. plastic, vital religiously	distorted, inadequate concepts	50	79	−05	−10	89
3. understanding	inward looking, self-seeing only	96	−04	−14	06	95
4. confused	on firm emotional foundation	−18	12	80	00	69
5. vital in thinking	conventional in thinking	84	21	05	42	93
6. gentle	severe, critical	88	26	12	−02	86
7. masculine dominating	leading, not driving	−93	−10	03	−02	88
8. likeable	irritating	69	49	−03	−08	72
9. devoted to a purpose	takes life as it comes	32	65	02	00	41
10. not satisfied with taking other people's thinking	does no individual thinking	63	23	20	53	77
11. can talk about philosophical ideas	not interested in philosophical ideas	73	57	08	26	94
12. broader outlook	limited outlook	85	32	00	41	99
13. stolid	high strung	−23	00	−71	−04	56
14. resentful	accepting	−97	03	02	02	94
15. don't see me as a person	sympathetic	−91	−03	−07	10	84

fused about the future and considering a fairly radical vocational change. Without going into the protocol in the same detail as the previous one, one is immediately struck by her concern about being accepted, about a *Weltanschauung*, and about emotional stability. If one were working with this woman clinically, he would be particularly interested in the relationship between the construct "masculine dominating–leading, not driving" and the rest of the factor on which it is loaded. Does she indeed see males as dominating, "set in ways," and generally unthinking? If so, how does this enter into her present difficulty? Does she see herself frustrated because this is essentially "a man's world"? These are just a few of the questions and hypotheses which a clinician would raise on the basis of this protocol.

Discussion

We have presented what we believe to be a useful and reliable approach to the "mapping" of the individual cognitive structure. In its present form it would be impractical except for research purposes. However, Kelly (3) is developing a nonparametric equivalent to factor analysis which would permit one to factor a protocol in about one hour's time. We would like in this section to sketch some of the research possibilities which we believe are opened up by this method.

Not losing sight of the fact that the RCRT is a sorting test akin to other concept-formation tests, one immediately wonders what the qualitative relationships might be between the constructs elicited by it and the concepts elicited by these tests. More specifically, since here we obtain a picture of the relationship between constructs, there are certain questions amenable to investigation for the first time. For example: What is the significance of the number of factors derived? Is there a relationship between number of factors derived and complexity of cognitive structure? If so, what are the behavioral correlates of complexity of cognitive structure? Does the number of constructs having significant loadings on a given factor have any significance? One might suspect that the larger the number of constructs significantly loaded on a given factor, the more important that dimension is for the individual. Does factorial structure change along developmental lines?

Aside from the content itself, there are certain implications for clinical research. Since the factors derived are to a good extent the result of functional relationships which the individual himself has imposed upon his constructs, one wonders at the extent to which he is aware of this functional relationship. As a corollary, one might be interested in the correlates of accuracy of awareness of functional relationships between constructs. Such relationships could be studied by having the individual attempt to sort his constructs into as many groups as there were factors, and determining the correspondence between his grouping and the factor grouping. Correspondence could be construed as one form of insight and might be expected to vary with personal adjustment, age, success or failure in psychotherapy, etc. The clinician, interested as he is in problems of interpersonal relationship, could easily investigate such areas as transference, identification, etc., by means of an inverse analysis of the same protocol using the role titles as the variates. Here again it would be interesting to study developmental trends.

Other research opportunities present themselves in the fields of social and industrial psychology where one might be interested in, among other things, the commonality of constructs among members of a given group, the relationship between nature and structure of cognitive structure and leadership ability, etc. The problem of communication could be approached through the study of the personal constructs of the communicators. From the standpoint of communication, this technique provides a means of reproducing each individual's personal coding scheme. One might study such problems as ability to receive and transmit information as a function of the content and structure of individual coding systems.

One last point should be made with respect to the problem of validity. In one sense there is no such thing as an invalid test, since all tests measure something, and it becomes a question of semantics whether the right name has been applied to it; the more important question as we see it is *pertinence*. The responses of an individual on the RCRT constitute a phenomenon; by means of factor analysis we impose a certain structure upon this phenomenon, and now the problem becomes one of the pertinence of this phenomenon and this structure. The entire body of this discussion has been devoted to the description of a program which we believe would yield an answer to this question. To the extent that the information yielded by the RCRT and its factor analysis is found to be functionally related to other psychological phenomena, the technique has pertinence. In this spirit we offer the technique to the psychological public.

Summary

A method is described whereby factor analytic techniques may be applied to the concepts formed by an individual on a sorting test called the Role Construct Repertory Test. Two cases are presented for illustrative purposes and the research implications of this method are discussed.

Received May 9, 1955.

References

1. Hunt, D. Studies in role construct repertory: Conceptual consistency. Unpublished master's thesis, The Ohio State Univer., 1951.
2. Kelly, G. A. *The psychology of personal constructs.* New York: Norton, 1955.
3. Kelly, G. A. The analysis of a person's psychological space. Unpublished manuscript, The Ohio State Univer., 1952.
4. Osgood, C. E. The nature and measurement of meaning. *Psychol. Bull.,* 1952, **49**, 197–237.
5. Wherry, R. J. A new iterative method for correcting erroneous communality estimates in factor analysis. *Psychometrika,* 1949, **14**, 231–241.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Efficiency of Attitudes, Fantasies, and Life History Data in Predicting Observed Behavior[1]

George W. Fairweather, Louis J. Moran, and
Robert B. Morton

VA Hospital, Houston, Texas

This study investigates the efficiency with which three frequently used sources of information, viz., attitudes, fantasies, and life histories, predict tuberculous patients' overt ward behavior.

In structured interviews, each of 140 patients responded to 132 life history items and 45 attitude items. In addition, ten TAT cards were administered to 39 of the 140 patients. Concurrently, all patients were rated by two nurses and two aides on 64 ward behavior items.

From this pool of items, three ward behavior scales, three attitude scales, and two life history scales were constructed. The three ward behavior scales and the three attitude scales concerned adjustment to (*a*) regulations, (*b*) personnel, and (*c*) peers. The two life history scales concerned developmental adjustment to (*a*) control and (*b*) peers. Each response to every item in all scales was given an a priori adaptation score so that the higher the adaptive significance of a response the higher the score. The total score for any scale was the sum of the item values comprising that scale.

The 39 TAT protocols were scored, by frequency count, on the following fantasy dimensions: (*a*) hostility, (*b*) clear or weak perception of the parent, (*c*) aspiration, and (*d*) inactivity. In addition, 33 of the 140 patients, who remained in the hospital 8 to 15 months after the initial ward behavior ratings, were again rated on the same three behavior scales.

A 12 by 12 tetrachoric correlation matrix was computed with the initial data. The matrix revealed that attitudes, life history data, and fantasy dimensions were not significantly related to the scores on the three ward behavior scales. However, the correlations among the three ward behavior scales were .66, .66, .76 ($p < .001$). Further, the follow-up ward behavior scores yielded the following product-moment correlations with the initial ward behavior scores: (*a*) adaptation to regulations .60 ($p < .01$), (*b*) adaptation to personnel .27 ($p > .05$), (*c*) adaptation to peers .59 ($p < .01$).

These results indicate that the most efficient predictor of one aspect of current behavior is another aspect of current behavior. For example, current adaptation to regulations predicts, with reasonable accuracy, the patients' current adaptation to personnel and peers. Further, the significant correlations between initial and follow-up behavior scales in the areas of adaptation to regulations and peers intimate that one aspect of behavior predicts, within moderate limits, subsequent similar behavior. More generically, the findings suggest that the closer in conceptual distance the independent variable is to the dependent variable, the more accurately will the latter be predicted from the former.

Brief Report
Received September 12, 1955.

---

[1] An extended report of this study and test materials may be obtained without charge from Robert B. Morton, Veterans Administration Hospital, Houston 31, Texas, or for a fee from the American Documentation Institute. To obtain it from the latter source, order Document No. 4692 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.75 for microfilm or $2.50 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Double Alternation: A Measure of Intelligence[1]

Allen Hodges
Southern Minnesota Mental Health Center

Since its formulation by Hunter (5), the double alternation problem has been recognized as a measure of an implicit process designated as reasoning. The subject, whether animal or human, must respond in terms of temporal relations under conditions where no differential sensory cues exist. Because of its adaptability, versions of the double alternation problem have been widely used in comparative studies of intelligence in mammals (1, 4, 5, 7, 8).

Four studies have been reported using the double alternation technique as a measure of intelligence with human subjects. With college students as subjects, Gellermann (2) reported correlations between solution of the problem and chronological age and mental age to be .28 ± .11 and .58 ± .09, respectively. Hunter and Bartlett (6), using children between the ages of 2 and 6, found a correlation of .86 ± .03 between a number of trials required to solve the problem and chronological age; a correlation of .81 ± .04 was found between a number of trials and mental age.

Stolurow and Pascal (10), using mental defectives as subjects, found double alternation performance to correlate .83 ± .03 with measured mental age, while no significant relationship was found with chronological age. Hodges (3), using public school children, reported biserial correlations of .78 ± .03 and .59 ± .05 between solution of the double alternation problem and CA and MA.

From these studies with human subjects, the relationship between measured intelligence and double alternation behavior is evident. The purpose of this paper is to equate measured mental age with performance on the double alternation problem. Because of the nonverbal qualities of double alternation behavior, a test of this behavior could be a valuable addition to the clinician's test battery.

Method and Procedure

Selection of Subjects

A stratified sample of elementary school children was selected.[2]

The following selective criteria were used: (*a*) Only male subjects are included. (*b*) All subjects were administered the Primary Mental Abilities Test (11). (*c*) Two hundred and forty subjects with PMA IQ's ranging from 80 to 120 were subdivided into six chronological age groups. (*d*) While subjects were not individually selected on the basis of socioeconomic status, all subjects were drawn from school districts in which family earning power, educational level of parents and type of housing were comparable.

Table 1 contains the chronological and mental age characteristics of the six age groups employed.

Table 1

Chronological and Mental Ages of the Subjects

Age group	N	Chronological age mean	σ	Mental age mean	σ
6 to 7	40	6.54	.31	6.21	.84
7 to 8	40	7.50	.28	7.59	.93
8 to 9	40	8.41	.34	8.15	.77
9 to 10	40	9.40	.39	9.25	.81
10 to 11	40	10.39	.36	10.41	.63
11 to 12	40	11.41	.24	11.42	.69

---

[1] The author is indebted to Drs. G. R. Pascal, E. O. Milton, and E. E. Cureton of the University of Tennessee for advice and guidance given in completing the dissertation upon which this article is based.

[2] All subjects were obtained from the Oak Ridge, Tennessee, Public Schools.

The Double Alternation Card Test

Each of the 240 subjects was administered the Double Alternation Card Test individually. The Card Test was introduced by J. McV. Hunt (9).[3] The only materials needed are five ordinary playing cards and a record sheet. Four of the five playing cards are black (either spades or clubs) while the fifth is red (heart or a diamond). The red card is designated as the goal card.

The five cards are arranged out of vision of the subject and placed face down on the examining table. The goal card is placed according to the double alternation sequence, twice on the left end and twice on the right end (LLRR). The subject is simply instructed: "Find the red card."

The complete sequence of LLRR or four complete presentations comprises one trial. The subject's errors are recorded as the subject seeks the goal card. The test is continued for thirty trials or until the subject successfully solves the problem by picking the goal card for two successive trials.

Six quantitative scores were derived from the Card Test.

1. Success or Failure. This is a dichotomy score of whether the subject obtains the criterion of two correct successive trials before a total of 30 trials is presented.

2. Verbalization or Nonverbalization Score. If the subject successfully solves the problem, this score is whether the subject can verbalize the answer. With young children, the problem is scored as "Verbalized" if the child states: "You put the red card two times here (points correctly) and two times here (points correctly)."

3. The Number of Trials Score. This score is the number of trials which the subject uses before obtaining the criterion of success.

4. The Number of Total Errors Score. This score is the total number of errors made by the subject before obtaining the solution, or the total errors made during the 30 trials.

5. The Errors Prior to End Concept Score. In observing subjects confronted with the double alternation problem, their first behavior appears to be trial and error. After locating the red card on the ends, the subject recognizes that the goal card is always placed in the end-most positions. From this time on, the majority of subjects will turn only the end cards. This point at which the *S* turns only the end cards for two successive trials is designated as the development of the concept of "endedness." The errors committed prior to achieving this concept comprise the Errors Prior to End Concept Score.

6. The Errors after End Concept Score. This score is the number of errors committed after formation of the end concept.

A seventh score is obtainable which is the total Time Score, or the time which elapses between the first presentation of the cards and the solution to the problem. Because the dexterity of the examiner in handling and distributing the cards contributes to the time involved in testing, the validity of such a score is questionable.

Table 2

Correlation Summary

Measures	Solution vs. failure 1	Verbalization vs. nonverbalization 2	Number of trials 3	Total number of errors 4	Errors prior to end concept 5	Errors after end concept 6	Mental age 7
1. Solution vs. failure							
2. Verbalization vs. nonverbalization	.800*						
3. Number of trials	−.680*	−.760*					
4. Total number of errors	−.633†	−.595†	.894†				
5. Errors prior to end concept	−.219†	−.276†	.344†	.765			
6. Errors after end concept	−.579†	−.468†	.702†	.136	.149		
7. Mental age	.597†	.536†	−.543†	−.743	−.222	−.487	

* Tetrachoric coefficient.
† Biserial coefficient.

---

[3] In a forthcoming article, Dr. Hunt will describe the Card Test in more detail.

Table 3

Mean Scores of the Six Double Alternation Measures for Each Mental Age Group

MA	N	Solution or failure	Verbalization or nonverbalization	Mean number of trials	Mean total number of errors	Mean errors prior to end concept	Mean errors after end concept
14 to 15	6	S	V	5.8	18.3	14.8	3.8
13 to 14	7	S	V	9.3	20.3	16.4	8.1
12 to 13	16	S	V	11.6	29.7	14.5	12.8
11 to 12	32	S	V	12.4	35.8	20.2	16.3
10 to 11	35	S	NV	18.2	55.9	32.2	24.4
9 to 10	37	S	NV	19.7	54.2	22.2	25.1
8 to 9	46	S	NV	21.5	61.9	36.7	31.6
7 to 8	36	S−/F	NV	21.7	58.7	24.7	34.2
6 to 7	21	F	NV	26.7	87.3	47.0	41.1
5 to 6	4	F	NV	30.0	99.3	39.2	60.0

Results

Table 2 contains the intercorrelation summary of the six scores on the Card Test and the criterion mental age.

Solution vs. Failure, Total Number of Errors, and Errors after End Concept Scores were selected as having the highest correlation with the criterion variable and lowest intercorrelations with the other variables. A multiple regression equation was computed by the Doolittle method using these three scores. The result was:

$$\text{Estimated mental age} = .24\ (X1) - .49\ (X2) - .09\ (X3) + C$$

Where: $X1$ = Solution vs. Failure Score (Solve = + .65, Fail = − .97)[4]

$X2$ = The square root of the Total Number of Errors Score[5]

$X3$ = The square root of the Errors after End Concept Score[5]

$C = 12.15$

This equation represents the best estimate of mental age obtainable for the three Card Test scores. The use of this equation is limited, however, because of the restricted range of ability which it encompasses; i.e., predicting mental ages below a twelve-year level. In addition, the need for cross validation is evident. When this equation is applied to a new group of subjects, considerable shrinkage may occur which would alter the predictive value of the multiple regression equation.

To increase the clinical usefulness of the Card Test, mean scores have been compiled which equate scores on the Card Test's six measures with mental age. These means are included in Table 3.

Summary and Conclusions

1. The purpose of this paper was to equate performance on the Double Alternation Card Test with measured mental age. Using 240 public school males ranging in CA from 6 to 12 years, mental age equivalents for six quantitative scores of the Card Test were computed.

2. A multiple regression equation predicting mental age was derived. The usage of this equation appears limited because of the restricted mental age range which can be predicted by means of it.

3. Mean scores for performance on six measures of the Card Test, equated with PMA mental ages, are included as very tentative norms.

4. Because of the nonverbal aspects of the Card Test plus the high correlations between

---

[4] To determine the scores to be assigned for Solution vs. Failure, the values corresponding to P, or the percentage of solution, and Q, which is the percentage of failure, for the mean deviation of one tail of a normal distribution was computed.

[5] The square roots of these scores are necessary because of a square root transformation previously applied to the data.

measures on this test and measured mental age, it is felt that this technique can be a valuable addition to the clinician's test battery, particularly in instances where the testee suffers language handicaps, aphasia, or speech disorders.

Received May 13, 1955.

References

1. Gellermann, L. W. The double alternation problem: I. The behavior of monkeys in a double alternation temporal maze. *J. genet. Psychol.,* 1931, 39, 50–72.
2. Gellermann, L. W. The double alternation problem: II. The behavior of children and human adults in a double alternation temporal maze. *J. genet. Psychol.,* 1931, 39, 197–226.
3. Hodges, A. A developmental study of symbolic behavior. *Child Develpm.,* 1954, 25, 277–280.
4. Hunter, W. S. The temporal maze and kinesthetic sensory processes in the white rat. *Psychobiology,* 1920, 2, 1–18.
5. Hunter, W. S. The behavior of raccoons in a double alternation temporal maze. *J. genet. Psychol.,* 1928, 35, 374–388.
6. Hunter, W. S., & Bartlett, S. C. Double alternation behavior in young children. *J. exp. Psychol.,* 1948, 38, 558–567.
7. Hunter, W. S., & Hall, B. E. Double alternation behavior of the white rat in a spatial maze. *J. comp. Psychol.,* 1941, 32, 253–266.
8. Hunter, W. S., & Nagge, J. W. The white rat and the double alternation temporal maze. *J. genet. Psychol.,* 1931, 39, 303–319.
9. Nichols, P. C., & Hunt, J. McV. A case of partial bilateral frontal lobectomy. *Amer. J. Psychiat.,* 1940, 96, 1063–1083.
10. Stolurow, L. M., & Pascal, G. R. Double alternation behavior in mental defectives. Paper read at Amer. Psychol. Ass., Penn. State Col., 1950.
11. Thurstone, L. L. *The Primary Mental Abilities Test.* Chicago: Science Research Associates, 1948.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Estimating the Full Scale Score on the Wechsler Adult Intelligence Scale from Scores on Four Subtests

Jerome E. Doppelt
The Psychological Corporation

Numerous articles have appeared in the literature on abbreviated forms of the Wechsler-Bellevue Scale. The short forms have usually been based on data from samples which were not at all typical of the normal population. In 1950, McNemar (2) reported a study on abbreviated Wechsler-Bellevue Scales which was based on the data Wechsler (3) reported for his standardization population. McNemar gave the correlations between total score on the Wechsler-Bellevue and the best pairs, triads, quartets, and quintets of subtests. He makes the point that "there can be no doubt about the greater dependability of the correlations based on Wechsler's group," that is, the standardization sample for the W-B Scale, than on attempts with less representative groups.

Since February 1955, the Wechsler Adult Intelligence Scale (WAIS), the revision of the Wechsler-Bellevue Scale, has been available. The WAIS consists of six verbal and five performance subtests, and like its predecessor it yields Verbal, Performance, and Full Scale IQ's according to age group. There are many instances when users of the WAIS would want a reasonably accurate estimate of a subject's IQ without giving all eleven subtests of the Scale. This study is an investigation of the effectiveness of a subgroup of tests in predicting the Full Scale Score, which is the sum of scores on the eleven tests.

Following McNemar's suggestion, the data used were those gathered in the national standardization of the WAIS and reported in the Manual (4). These data are based on samples reasonably representative of the population of the United States.[1]

The decision as to the number and type of subtests to be included in the predictor group is, to some extent, arbitrary. A compromise must be made between economy of time and effort and accuracy of prediction. It was decided to select the group of four subtests which correlates highest with Full Scale Score. Although prediction of the Full Scale Score was the goal, it was felt that the best approach would be to select the two verbal subtests which are most highly correlated with total Verbal Score and the two performance measures which are the best predictors of the total Performance Score.

There are fifteen possible pairs of the six verbal subtests of the WAIS and the correlation of each pair with the Verbal Score (total of scores on the six tests) was computed for three different age groups: 18–19, 25–34, and 45–54. The coefficients are reported, in order of magnitude, in Table 1. For the ten pairs of performance tests the correlation coefficients between each pair and Performance Score are reported for the same age groups in Table 2.[2]

Tables 1 and 2 show that many of the pairs of tests are highly correlated with the corresponding total scores and the differences

---

[1] Full description of the sampling procedure used in the collection of the data is given in the Manual (4, pp. 5–11).

[2] The basic data for these calculations are given in Tables 7–9 of the WAIS Manual (4). For greater accuracy in computation the intertest coefficients to four decimal places were used. These data were available in the files of the publisher of the WAIS.

Table 1

Correlations Between Two Verbal Tests and Total Verbal Score

Ages 18–19		Ages 25–34		Ages 45–54	
Tests	r	Tests	r	Tests	r
Inf., Voc.	.940	Inf., Voc.	.940	Arith., Voc.	.948
Arith., Voc.	.938	Inf., Sim.	.936	Inf., Comp.	.941
Sim., Voc.	.938	Arith., Voc.	.934	Inf., Voc.	.940
Inf., Sim.	.936	Sim., Voc.	.924	Inf., Arith.	.939
Comp., Voc.	.933	Inf., Comp.	.921	Comp., Voc.	.935
Comp., Sim.	.931	Comp., Arith.	.914	Arith., Sim.	.933
Inf., Arith.	.930	Inf., Arith.	.913	Comp., Arith.	.930
Arith., Sim.	.929	Comp., Voc.	.912	Sim., Voc.	.930
Inf., Comp.	.925	Dig. Sp., Voc.	.912	Inf., Sim.	.930
Inf., Dig. Sp.	.925	Arith., Sim.	.911	Dig. Sp., Voc.	.928
Sim., Dig. Sp.	.914	Comp., Sim.	.908	Comp., Sim.	.920
Dig. Sp., Voc.	.912	Inf., Dig. Sp.	.905	Inf., Dig. Sp.	.918
Comp., Arith.	.911	Comp., Dig. Sp.	.894	Comp., Dig. Sp.	.909
Comp., Dig. Sp.	.906	Sim., Dig. Sp.	.891	Sim., Dig. Sp.	.903
Arith., Dig. Sp.	.881	Arith., Dig. Sp.	.853	Arith., Dig. Sp.	.865

between the coefficients for successive pairs in any age group are very small. The coefficients are, of course, influenced by the fact that each pair of tests is included in the total with which it is correlated. However, the purpose here is to predict the total on the basis of part of the total and consequently this "spurious" nature of the coefficients is of no concern.

In Table 1, the Information and Vocabulary subtests constitute the best pair for two of the age groups and the third best pair for one group. The pair which includes Arithmetic and Vocabulary is in second, third, and first position, respectively, for the three age groups. Since the difference between these "best" pairs is negligible, the choice of the pair to use in a regression equation was open to other than purely statistical considerations. Arithmetic and Vocabulary were selected as the pair of verbal subtests to use because they represent a variety of content materials. In particular, the inclusion of the arithmetic content provides a desirable range of situations for subjective observations in the testing session without loss of statistical accuracy.

The Block Design and Picture Arrangement pair is most highly correlated with Perform-

Table 2

Correlations Between Two Performance Tests and Total Performance Score

Ages 18–19		Ages 25–34		Ages 45–54	
Tests	r	Tests	r	Tests	r
Bl. Des., P. Arr.	.939	Bl. Des., P. Arr.	.917	Bl. Des., P. Arr.	.926
P. Arr., Obj. Assem.	.920	Dig. Sym., Bl. Des.	.917	P. Arr., Obj. Assem.	.926
P. Comp., Bl. Des.	.920	P. Comp., Bl. Des.	.914	P. Comp., Obj. Assem.	.920
Bl. Des., Obj. Assem.	.920	P. Comp., Obj. Assem.	.909	P. Comp., Bl. Des.	.919
Dig. Sym., Bl. Des.	.918	P. Arr., Obj. Assem.	.906	Dig. Sym., Obj. Assem.	.919
P. Comp., Obj. Assem.	.916	P. Comp., P. Arr.	.906	Dig. Sym., Bl. Des.	.915
P. Comp., P. Arr.	.913	Bl. Des., Obj. Assem.	.902	Dig. Sym., P. Comp.	.914
Dig. Sym., P. Comp.	.912	Dig. Sym., P. Comp.	.900	P. Comp., P. Arr.	.905
Dig. Sym., Obj. Assem.	.906	Dig. Sym., Obj. Assem.	.897	Dig. Sym., P. Arr.	.887
Dig. Sym., P. Arr.	.887	Dig. Sym., P. Arr.	.880	Bl. Des., Obj. Assem.	.886

ance Score in all three age groups. This pair was selected to represent the performance tests.

The sum of the scaled scores on the four selected subtests—Arithmetic, Vocabulary, Block Design, and Picture Arrangement—was correlated with the Full Scale Score for each of the three age groups shown in Tables 1 and 2. The computation was done by the following formula with all the terms available from the tables of intercorrelations:

$$r_{(a+b+c+d)T} = \frac{r_{aT}\sigma_a + r_{bT}\sigma_b + r_{cT}\sigma_c + r_{dT}\sigma_d}{\sigma_{(a+b+c+d)}}.$$

The subscripts a, b, c, and d identify the four subtests. T identifies the Full Scale Score. The denominator was computed as follows:

$$\sigma_{(a+b+c+d)} = \sqrt{\sum_{i=a}^{d}\sigma_i^2 + \sum_{i=a}^{d}\sum_{j=a}^{d} r_{ij}\sigma_i\sigma_j} \qquad (i \neq j).$$

In the process of obtaining norms on the WAIS for older people, four "old-age" groups had been obtained and the intercorrelations among the WAIS tests had been computed (1). For each of these four groups the correlation between the selected four tests and the Full Scale Score was determined. The resulting coefficients and the regression equations for predicting Full Scale Score from the sum of four tests are shown in Table 3 for the three age groups in the national standardization and for the four old-age groups.

It may be noted from Table 3 that the slopes of the regression lines are very similar in the seven age groups. Indeed, it seemed practical to use 2.5 as the coefficient of the predicting variable in each of the equations and to adjust the constant term according to age of the individual. In this way a regression equation could be determined for each age group, including those for which tables of intercorrelations of the tests had not been prepared. This table, with the constants given as whole numbers, is presented as Table 4. Thus, an examiner who obtains the scaled scores on Arithmetic, Vocabulary, Block Design, and Picture Arrangement may estimate the Full Scale Score by multiplying the sum of the four scores by 2.5 and adding a constant in accordance with the subject's age. An estimate of the Full Scale IQ may then be obtained by entering the IQ tables given in the WAIS Manual.

Table 3

Correlation of Sum of Four Subtests with Full Scale Score and Regression Equation for Predicting Full Scale Score

Age group	N	r	Regression equation (x_1=Full Scale Score; x_2=sum of 4 Tests)
18–19	200	.960	$x_1 = 2.5x_2 + 9.7$
25–34	300	.954	$x_1 = 2.4x_2 + 13.0$
45–54	300	.958	$x_1 = 2.5x_2 + 7.5$
60–64	101	.968	$x_1 = 2.3x_2 + 12.2$
65–69	86	.963	$x_1 = 2.3x_2 + 10.7$
70–74	80	.957	$x_1 = 2.4x_2 + 7.9$
75 and over	85	.962	$x_1 = 2.6x_2 + 0.2$

Table 4

Estimation of Full Scale Score from Sum of Scores on Arithmetic, Vocabulary, Block Design, and Picture Arrangement

Age group	Full Scale Score= 2.5 S*+
16–17	10
18–19	10
20–24	10
25–34	10
35–44	9
45–54	8
55–64	7
65–69	5
70–74	5
75 and over	4

* S = sum of scaled scores on 4 tests (use Table 17 of WAIS Manual).

In making any prediction one should always have some idea of the error involved. In this study the standard deviation of Full Scale Scores is about 25 and the correlation coefficient between the sum of the four selected tests and Full Scale Score is approximately .96. Consequently the standard error of estimate in predicting Full Scale Score from the four tests is about 7 scaled score points. (This is equivalent to 4.2 IQ points.) Thus an estimated Full Scale Score would be within 7 scaled score points of the actual score about 68 out of 100 times.

As a check on the equations given in Table 4, distributions of the actual differences between the obtained and estimated Full Scale

Table 5

Distribution of Differences Between Full Scale Score and Estimated Full Scale Score

	Ages 20–24		Ages 35–44	
Difference*	*f*	%	*f*	%
Over 21	1	0.5		
15–21	3	1.5	2	0.7
8–14	25	12.5	39	13.0
0–7	70	35.0	115	38.3
−7–−1	72	36.0	98	32.7
−14–−8	24	12.0	37	12.3
More than −14	5	2.5	9	3.0
N	200		300	
Mean	−0.1		−0.5	

* Full Scale Score minus the estimated Full Scale Score obtained from Table 4.

Scores were prepared for age groups 20–24 and 35–44. These two groups had not been used in the correlation analysis which led to the regression equations although they were used in the determination of age-group constants.

The distribution of differences, grouped in terms of the suggested standard error of estimate of 7 points, is shown in Table 5. It may be noted that 71 per cent of the differences are contained within one standard error (± 7 points) and 96 per cent are within two standard errors (± 14 points) in each age group.

Summary

An abbreviated form of the Wechsler Adult Intelligence Scale was determined by selecting the two best predictors of the total Verbal Score and the two best predictors of the total Performance Score. The four tests were Arithmetic, Vocabulary, Block Design, and Picture Arrangement. The selection of tests was based on the data obtained in the national standardization of the WAIS.

The correlation coefficients between the sum of scaled scores on the four tests and the Full Scale Score varied between .95 and .96 for the seven age groups studied. A simplified regression equation for predicting Full Scale Score was presented in which the sum of scaled scores on the four selected tests is multiplied by 2.5 and a constant is added depending upon the subject's age.

The selection of subtests made in this study permits an estimate of Full Scale Score after 35–40 minutes of testing. The standard error of estimate of Full Scale Scores thus obtained is about 7 standard score points.

Received November 15, 1955.
Early Publication.

References

1. Doppelt, J. E., & Wallace, W. L. Standardization of the Wechsler Adult Intelligence Scale for older persons. *J. abnorm. soc. Psychol.*, 1955, **51**, 312–330.
2. McNemar, Q. On abbreviated Wechsler-Bellevue Scales. *J. consult. Psychol.*, 1950, **14**, 79–81.
3. Wechsler, D. *The measurement of adult intelligence.* (3rd Ed.) Baltimore: Williams & Wilkins, 1944.
4. Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale.* New York: Psychological Corp., 1955.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

The Wechsler-Bellevue and Psychiatric Diagnosis: A Factor Analytic Approach

George H. Frank[1]
Florida State University[2]

Although considerable research has been conducted to determine the efficacy of the Wechsler-Bellevue Intelligence Test (5) in predicting psychiatric diagnoses (2, 3), the results have been disappointing. By and large, the hypothesis that the W-B can elicit more than intellectual factors has not been empirically validated.

The usual method of testing this hypothesis is to select two distinct samples, either two clinical groups, or a clinical and a so-called "normal" population, and compare the group subtest performance by t or r. Through the use of such techniques the researcher is afforded only a gross measure of disparity or similarity without any further information as to the reasons for the results.

An interesting innovation occurred to the writer which might provide a measure of comparability but, further, offer possible rationale for the similarity. It was reasoned thus: if one were to take the subtest scores of individuals from several diagnostic groups and throw them into one correlation matrix and factor-analyze the matrix, the subjects with similar subtest patterns would be grouped, and the reasons for their groupings could be ascertained from other identifying data.

Method

In a previous research (1), analyses of variance of the subtest scores of subjects subsumed under thirty different diagnostic categories were performed. This analysis demonstrated that only nine of these constituted homogenous groupings of subtest scores, while the 21 others proved to be statistically unrelated.

A further check on the homogeneity of these nine groupings of W-B subtest scores was made by computing *W*'s (a method of multiple-rank correlation) for each of the nine groups. The *W*'s were high (above .90) and significant. Then the subtest scores of the sixty subjects from these nine different psychiatric groups were intercorrelated, one with the other, by the Spearman method of rank correlation, and the matrix analyzed by Thurstone's centroid method of factor analysis.

Since the data originally came from Rapaport's work (4), it was a simple task to turn to the data for the identifying characteristics of the factor loadings. Thus those high in a given factor were compared on the basis of their diagnosis, age, sex, education, verbal, performance, and Full Scale IQ's to determine the communality.

Results

Because of the cumbersomeness of a 60 × 60 correlation table, the original matrix is not presented here. However, Table 1 presents the factor loadings of the sixty subjects on the two factors isolated by the analysis of the correlation matrix.

The question now arises as to the meaning of the factor loadings. The immediate conclusion proffered by the data is that psychiatric diagnosis is not the common element of either one of the factor loadings. In fact the

---

[1] Based in part on a paper read at the Southeastern Psychological Association, in Atlanta, on May 23, 1955.

[2] Now at Topeka State Hospital, but the writer would like to acknowledge the assistance of Prof. Bob McGinnis, now of the University of Wisconsin, then Director of the Sociological Research Laboratory of Florida State University.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Anxiety Level and Pursuitmeter Performance[1]

Ruth G. Matarazzo and Joseph D. Matarazzo
Massachusetts General Hospital and Harvard Medical School

The *S*s were 80 white, male, VA inpatients ranging in age from 18 to 37. The measure of anxiety employed was the Taylor Manifest Anxiety Scale. Five anxiety groups were used, Taylor raw scores of 1 to 8 being given an anxiety rating of 1, 9 to 16 a rating of 2, etc. Group 5 contained individuals with Taylor scores from 33 to 44. The number of *S*s within each group from 1 to 5 were 7, 17, 16, 21, and 19, respectively. The five anxiety groups were equated for age, education, and total Wechsler-Bellevue IQ. Also there were no significant group differences in performance on any of the eleven W-B subtests, as measured by F tests and epsilon correlations. The finding of no significant differences in W-B scores is important since both Grice and Kerrick and our own group have reported moderate but significant negative correlations (ranging from $-.20$ to $-.40$) between the Taylor Scale and some measures of "intelligence." For the present data, the value of the correlation (epsilon) was not significant.

The learning task was a complex, double-disk pursuitmeter loaned by D. Lewis. For each S there were 20 trials of 20 seconds each followed by 40 seconds of rest. Two measures of learning were used: *gain* (the difference between each S's average score on the last two trials and his average score on trials 1 and 2), and *total time on target*.

Results indicate no statistical relationship between either of these two learning measures and Taylor anxiety level. Despite this lack of significance as shown by F tests (.29 and .00) and correlation measures (epsilon values of .04 and .22), there was a trend for the middle anxiety groups (2 and 3) to be superior learners. This latter finding, while of no interest in itself, takes on some slight significance in view of the finding of a similar curvilinear relationship in two previous studies from our laboratory, and two unpublished studies of Spence and Taylor from the Iowa laboratory.

The present negative results are of interest in view of the findings of both Grice and Kerrick, and our own previous work on the relationship between "intelligence" and score on the Taylor scale. Furthermore the present findings do not support our earlier proposed hypothesis that *timed* tests might be more sensitive to variations in Taylor anxiety level than are other tests. Nor do the present findings support our earlier belief that the trend toward a curvilinear relationship, found with a digit-symbol learning task, was an example of the type of case described by Webb and Lemmon, in which the interpretation of a nonsignificant F test should be qualified.

Brief Report.

Received October 28, 1955.

---

[1] An extended report of this study may be obtained without charge from Ruth Matarazzo, Massachusetts General Hospital, Boston 14, Massachusetts or for a fee from the American Documentation Institute. Order Document No. 4748 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.75 for microfilm or $2.50 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

A Validation Study of the Bender-Gestalt[1]

Benjamin Mehlman and Edward Vatovec
Kent State University

In recent years, increasing attention has been given to the application of test procedures to the problems of organic brain pathology. If, as implicitly and explicitly indicated (3, 7, 9, 11), there are dimensions of behavior that are peculiar to the organic, then tests may be evolved which can distinguish the organic. The purpose of this study is to determine whether the Bender Gestalt Test reliably differentiates organic from functional psychotic institutionalized individuals.

Before the introduction of the Bender Gestalt Test, perception and reproduction of design by psychotic patients had received only scant attention by other authors (11). In 1938, Bender (3, 4) assembled nine Gestalt figures derived from Wertheimer's configurations into what is now known as the Bender Visual Motor Gestalt Test. This test was built on the premise that the reproduction of test design is a visual-motor act in which distortions of original pattern indicate malfunctioning due to neural injury, variations in intellectual levels, or maladjustment in the emotional field of the perceiving subject. Schilder (14), in the preface to Bender's monograph of the test (3), states that "Personal experience has taught me that the clinical value of the test is very great. It may allow a differential diagnosis between organic deterioration, so called functional mental disease and malingering."

Several studies with the Bender Gestalt have failed to differentiate functionally ill from nonfunctionally ill, and between groups of functionally ill (1, 5, 6, 13). However, Irion and Pascal (15, p. 148) and Hobbs (10) report reliable differentiations between psychiatric patients and nonpatients. Preliminary observations by Barkley (2) with plastic relief reproductions suggest differentiation between organic and nonorganic.

Halpern (8), writing about the Bender Gestalt in a text on projective techniques, reports that where the test has been given before and after brain injury or other traumatic experiences, the disturbance in personality found expression through the resulting distortions of the *gestalten*. Woltmann (16) summarizes the work reported by Bender on organic brain diseases in her monograph and says it proves beyond doubt that organic brain disease does interfere with the visual-motor gestalt functions.

These studies, with their variant conclusions as to the worth of the Bender Gestalt Test, have led to the current attempt to clarify the worth of this test. This study seeks to determine whether the Bender Gestalt reliably differentiates organic from functional psychotic institutionalized individuals.

Method

Psychotic patients from Massillon State Hospital, Massillon, Ohio, were used for this study. The name, sex, age, diagnosis, and the time spent in mental hospitals were recorded individually for each of the 3,216 patients at the hospital on separate file cards. These cards were then separated into four groups: one which was made up from male patients who carried an official hospital diagnosis of functional psychoses; another of female functional patients; one of male organic patients and the last of female organic patients. Eliminated from these groups were patients who carried a diagnosis of mental deficiency. The groups were then arranged in order of chronological age. It was possible individually to match 118 functionally psychotic male patients with 118 organically psychotic male patients of the same age and of not more than one year's difference in regard to time spent in the hospital. This same procedure was followed in choosing the female population of the study. There were 102 paired female patients following this procedure.

---

[1] Based on a master's thesis submitted by the junior author.

Eliminated from this study were patients who had undergone any form of psychosurgery. Also eliminated were any patients who had undergone any form of convulsive treatment within the previous six months. Kalinowsky and Hock (12) summarized studies made in this field and came to the conclusion that although there is no actual impairment of cerebral functioning following a course of convulsive therapy, there is some question as to how long after treatments have been discontinued these functionings return to normal. The six-month period was here considered a long enough time for the effect of the treatment to disappear. Also eliminated from this study were patients who refused to take the test, were unable to comprehend the directions, had markedly poor eyesight or other obvious or known physical handicaps which interfered with ability to take the test.

The elimination of patients for the above reasons reduced the 118 paired males to 34 matched male patients, and the 102 paired female patients to 21 matched female patients. After the protocols had been collected, two sets of 25 pairs of protocols each were selected in unbiased fashion from among the pairs of protocols available.

The design of this study called for statistically better than chance differentiation by experts between organically and functionally ill from the set of 25 matched pairs of protocols. To give this task to individuals only vaguely or incompletely familiar with the Bender Gestalt Test as a clinical instrument would have loaded the data against finding positive results. Sending protocols to experts, defined as individuals who had published accounts of the clinical research use of the test, and who were enthusiastic about the test, increased the likelihood of securing positive results. There were eight such individuals located by survey of the literature. One refused to participate and several felt unable to participate, for reason of time or the like. Four accepted invitations to participate, but only three returned the materials in completed form. Each authority received 25 paired protocols, each protocol containing only the code letter "A" or "B," with instructions to place either letter under the organic or functional column on the answer sheet.

Results

In terms of our earlier expectancies, that there would be approximately eight judges participating, two sets of 25 pairs of matched protocols had been arranged. In terms of the completed data, two judges used one set of 25 and the other judge used the remaining set. Since the placing of a pair of protocols into one or another of the sets was a chance affair, there is no reason to believe that the sets are not equivalent in any respect. Set 1 had 15 pairs of males, 10 pairs of females, and Set 2 had 16 pairs of males, 9 pairs of females.

On the basis of chance alone we could expect each judge to get 12.5 choices correct. Judge A secured 11 successes out of the 25 choices in Set 1. Since this represents fewer successes than the Mean, and since we are here concerned only with a one-tailed test, this represents a failure on the part of the Bender Gestalt, as utilized by this expert, to discriminate successfully between the organically and functionally ill.

Judge B secured 17 correct choices of the possible 25 in Set 2. Using a one-tailed test, this represents successful choice at the .035 level of confidence. In other words, there are only 3.5 chances in 100 that Judge B could have been right this number of times on the basis of chance alone.

Judge C secured 18 correct choices of the possible 25 in Set 1. Using a one-tailed test, this represents successful choice at the .014 level of confidence. In other words, there are only 1.4 chances in 100 that Judge C could have been right this number of times on the basis of chance alone.

In summary then, we have found that two of our judges attained or closely approached better than chance expectancy success in discriminating the organics from the functionally psychotic through the Bender Gestalt, and the third judge failed to have more than chance success.

Discussion

Although the Bender Gestalt had been in existence since 1938, it was only during World War II, when there was a felt need for a reliable test to determine organic brain pathology, that the Bender Visual Motor Gestalt Test became the subject for investigation. In looking over the studies on the Bender Gestalt Test we can find no rigorous study of the Bender Gestalt's ability to differentiate between functional psychotic patients and organically psychotic patients.

The current study suggests that while this instrument is not used equally well by different experts, at least with respect to the highly focused task of this study, in the hands of some it can reliably distinguish organic from functional psychotic. While variability in skill in utilizing such an instrument is no new finding, such heterogeneity as found here among authorities so highly select is surprising. This is particularly so when we consider that the differentiations were to be made from among patients not selected as differential diagnostic problems; i.e., these protocols were taken from patients who, to the best of our knowledge, are typical of patients in general from their respective groups.

Very much to the point too is the fact that even the most successful of our experts was wrong in 7 of 25 choices. The statistical significance should not lead to complacency about the Bender Gestalt. This is particularly true when we recall that the task set for our experts here was far easier than the task the clinician meets in his daily routine; i.e., here the judge, having diagnosed one protocol, had also diagnosed the other. In addition, our judges had a much more gross kind of labeling to do than is generally encountered in clinical practice. That there is some basis for the enthusiasm expressed in the literature has been demonstrated here. But the need for greater refinement in the use of the Bender Gestalt is equally apparent.

Summary and Conclusions

This study was undertaken to determine whether the Bender Visual Motor Gestalt Test reliably differentiates organic from functional institutionalized individuals.

Twenty-five paired protocols were submitted to three authorities of the Bender Gestalt Test for blind analysis to determine whether the experts were able to pick out the organic and functional psychotic protocol from the matched pairs. Two judges attained or closely approached better than chance success and another had no more than chance success. These results suggest a surprising heterogeneity of skill in the use of Bender Gestalt, even among experts. These results suggest many mistaken diagnoses, even by the best judge. Further refinement in the use of the Bender Gestalt is indicated, but the utilization of the Bender Gestalt alone in clinical practice is not warranted.

Received June 14, 1955.

References

1. Addington, M. C. A note on the Pascal and Suttell scoring system of the Bender Gestalt test. *J. clin. Psychol.*, 1952, **8**, 312–313.
2. Barkley, B. J. A note on the development of the Western Reserve Hapto-kinesthetic Gestalt test. *J. clin. Psychol.*, 1949, **5**, 179–180.
3. Bender, Lauretta. A visual motor Gestalt test and its clinical uses. *Res. Monogr. Amer. Orthopsychiat. Ass.*, 1938, No. 3.
4. Bender, Lauretta. Instructions for the use of the visual motor Gestalt test. *Amer. Orthopsychiat. Ass.*, 1946.
5. Billingslea, F. Y. The Bender Gestalt: An objective scoring method and validating data. *J. clin. Psychol.*, 1948, **4**, 1–27.
6. Glueck, G. Psicipatologia della percezione della forma. Principi della forma nei disegue copiata di malati mentali. *Arch. Psicol. Neurol. Psichiat.*, 1940, **1**, 603–664.
7. Goldstein, K. *The organism*. New York: American Book Co., 1939.
8. Halpern, Florence. The Bender Gestalt Test. In H. H. Anderson & Gladys L. Anderson (Eds.), *An introduction to projective techniques*. New York: Prentice-Hall, 1951.
9. Harrower, Molly R. Changes in figure-ground perception in patients with cortical lesions. *Brit. J. Psychol.*, 1939, **30**, 47–51.

10. Hobbs, N. Psychological research and services in an Army Air Forces convalescent hospital. Unpublished doctor's dissertation, Ohio State Univer., 1946.
11. Hunt, J. McV. (Ed.) *Personality and the behavior disorders.* New York: Ronald Press, 1944.
12. Kalinowsky, L. B., & Hock, P. *Shock treatment, psychosurgery and other somatic procedures in psychiatry.* New York: Grune & Stratton, 1952.
13. Pascal, G. R., & Suttell, Barbara J. *The Bender Gestalt Test; Quantification and validity for adults.* New York: Grune & Stratton, 1951.
14. Schilder, P. Psychic disturbances after head injuries. *Amer. J. Psychiat.*, 1934, 91, 155–188.
15. Wischner, G. J., Rotter, J. B., & Gillman, R. D. Projective techniques. In S. W. Bijou (Ed.), *The psychological program in AAF Convalescent Hospitals.* AAF Aviation Psychology Research Reports No. 15. Washington: Government Printing Office, 1947. Pp. 144–157.
16. Woltmann, A. G. The Bender Visual-Motor Gestalt Test. In L. E. Abt & L. Bellak (Eds.), *Projective psychology.* New York: Knopf, 1950. Pp. 322–356.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

Ranking Methodology: the Development of a Judgmental Criterion with Clinical Case Histories

A. W. Bendig
University of Pittsburgh

Although the ranking method has been used to quantify clinical judgments (4, 10), little is known of the parameters affecting the reliability and validity of judgments measured by this method. Characteristics of the stimuli, scale, and judges have been shown to influence the reliability of rating scale judgments (2, 3, 6), but similar research on the ranking method is nonexistent.

The present study was concerned with the problem of using the ranking method to establish a criterion ordering of clinical material based upon the judgments of experts. The central problem in establishing such a judgmental criterion is lack of agreement among the experts. A relatively low average intercorrelation among judges does not necessarily imply individual judge unreliability, but that the experts may be using different bases in making judgments of clinical material. If two clinical psychologists rate or rank clinical material using similar bases of judgment we would expect them to show high intercorrelation, but if they look at different aspects of the material their intercorrelation might be expected to be disappointingly low. As an example, Bendig (1) found that for esthetic judgments interjudge reliability (average intercorrelation among judges) is not a simple rectilinear function of intrajudge reliability (retest correlation for single judges).

A statistical method that can be used to test whether clinicians are using different bases of judgment and also to purify a judgmental criterion is inverse factor analysis. Factor-analyzing a matrix of intercorrelated judgments from clinicians would reveal whether a single general "agreement" factor can explain the obtained intercorrelations or whether a series of independent group factors is necessary. In the latter case the group factors would offer a means of categorizing the clinicians as to those with similar and dissimilar bases of judgment. Also the clinical material could be made more factorially homogeneous by eliminating material which clinicians using dissimilar bases disagreed upon and retaining only material that they judge similarly, regardless of their basis of judgment. However, the "purified" material should be validated by having other judges similarly rank the "purified" set of clinical material.

Procedure

Subjects. The 50 judges consisted of seven clinical psychologists and 43 undergraduate students in two sections of a course in mental hygiene. All the clinical psychologists had received the Ph.D. degree and had worked for a minimum of one year in clinical practice with children and adolescents. The undergraduate *S*s had a minimum of three semester hours in psychology with many having more. The Ph.D. *S*s were contacted individually and asked to read and rank for adjustment level 15 abstracted case histories. They were instructed to use their own clinical definition of adjustment, but it was suggested that they use as their reference point the behavior of the average child in a normal environment. The undergraduate *S*s were contacted in class and a booklet containing either 10 or 15 of the same case histories was given to each. A sheet of instructions similar to those given the Ph.D. *S*s was attached to the booklet. The undergraduate *S*s read and ranked the cases outside of the class and returned the booklet at a subsequent class period.

Stimuli. The 15 case histories were abstracted from longer and more complete cases published in a number of texts on clinical child psychology and psychiatry.[1] A standard outline was used with each case and the information was abstracted under six areas that previous interviews with clinical child psychologists had indicated were most useful and necessary. Each case abstract was approximately one page of single-spaced typing in length. Of the 15 cases selected, 10 were boys and 5 were girls. The range in age of the cases was from 7 to 16 years old with the median at 12 years.

Analysis. The rankings of the seven clinical psychologist judges were intercorrelated by the rank-difference method and the resulting matrix of correlations subjected to a centroid factor analysis. The obtained orthogonal factors were analytically rotated to approximate simple structure. Each judge's ranks of the 15 cases were weighted by his factor loadings on the rotated orthogonal factors and a mean weighted factor score obtained for each of the cases on each factor. The five cases whose factor scores showed the greatest intracase discrepancy were then selected for elimination. Booklets containing the original 15 cases or the 10 cases after the above elimination process were distributed to the undergraduate student judges for their adjustment ranking judgments. The average intercorrelation among the student *S*s in the two judgmental groups (those ranking 10 cases and those ranking 15) and their average correlation with the ranks of the same cases assigned by the clinical psychologists were computed by the formulas given by Lyerly (9). The differences between these average correlations from the two groups of student *S*s were tested for significance.

Results

The average intercorrelation among the clinical psychologists ranking 15 case histories was .51 and ranged from .04 to .83. The centroid factor analysis of the 21 intercorrelations yielded three orthogonal factors and the analysis was replicated three times to stabilize the communality estimates. After the third iteration the mean absolute difference between the estimated and obtained communalities was .01. The median absolute residual correlation after extraction of the third factor was .03.[2] The original factor loadings were analytically rotated to approximate simple structure and the rotated factor loadings are found in Table 1. It can be seen that there was no general "agreement" factor among the judges, but the three group factors obtained accounted for 80 per cent of the intrajudge variability in judgment. Five of the judges showed large loadings on factor B which accounted for 39 per cent of judgmental variability, while six of the judges had significant loadings on factor A which measured another 29 per cent of intrajudge variance. Factor C was bipolar and comprised the remaining 12 per cent of the variation in judgments, with two judges on the high positive end of this dimension and one judge on the negative end. Judgmental communality as estimated by these three factors varied from 54 to 93 per cent for the seven clinical psychologists.

Table 1

Summary of Centroid Factor Analysis of Interjudge Correlations of Clinical Psychologists

Clinical Psychologist	Rotated Factor Loadings			
	A	B	C	h^2
A	01	87	00	76
B	79	54	00	92
C	73	10	02	54
D	38	43	75	89
E	33	90	08	93
F	68	14	43	67
G	39	83	−30	93
Percentage of Intrajudge Variance	29	39	12	80

---

[1] The author's appreciation is extended to Miss Phyllis Black for the abstracting of the case histories.

[2] A table giving the original and residual intercorrelation between the clinical psychologists, the original factor loadings, and the transformation matrix has been deposited with the American Documentation Institute. Order Document No. 4744 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Each clinical psychologist's rank judgments were then weighted by his factor loadings given in Table 1 and a mean weighted score on each of the three weighted scores was obtained for each case and the five case histories with the largest interfactor variability selected for elimination. New rank-difference correlations between the clinical psychologists for the remaining 10 cases were computed and it was evident that only one general factor could account for the matrix of interjudge correlations. The mean correlation between these judges on these 10 case histories was .69 and Kendall's coefficient of concordance (8) was .74. The case sums of ranks for either the 10 or the 15 case histories assigned by the clinical psychologists were used to establish criterion rank ordering of the cases. Applying the Spearman-Brown formula to the average interjudge correlations gave a group reliability of .88 for the 15 cases and a group reliability of .94 for the summed ranks of the 10 case histories.

Booklets containing a face sheet of instructions plus either the original 15 case histories or the above selected judgmentally homogeneous 10 cases were assembled and distributed to 43 students in two undergraduate sections of a mental hygiene course. These *S*s ranked the 10 or 15 cases for adjustment level and their assigned ranks were summed for each case. Average interjudge correlations among the undergraduate *S*s within each of the two groups and their average correlation with the criterion rankings of the same cases were computed by the formulas given by Lyerly (9). The differences between the two undergraduate groups of *S*s were tested for significance by the methods outlined by Edwards (5, p. 136) which Hartley (7) has shown to be applicable to rank correlations. The summed ranks given by the undergraduate *S*s to the cases were used to rank the cases and these ranks correlated with the criterion ranking of the clinical psychologists by the usual rank-difference method. These results can be found in Table 2. Both the average intercorrelation among the undergraduate *S*s ranking the 10 homogeneous cases and their average correlation with the criterion rankings of the clinical psychologists were significantly larger at the .01 level. The average interjudge correlations were .51 and .28 for the groups ranking 10 or 15 cases ($t = 2.68$) while their average correlations with the criterion rankings were .70 and .47 ($t = 3.37$). The total ranks of the 10-case group showed almost a perfect correspondence to the total ranks given by the clinical psychologists ($rho = .97$), while the total ranks of the 15-case group was lower ($rho = .76$).

Discussion

The results of the present study agree with other reports indicating that clinical psychologists as a group show a moderate average intercorrelation in their judgments, but that the variability in agreement level among various pairs of experts is extremely large. For example, the average intercorrelation among our seven clinical psychologists was .51, but single correlation coefficients ranged from .83 to .04. The factor analysis of these intercorrelations indicated the absence of a general factor of agreement and the presence of several group factors. We suggest that a categorization of judges on the basis of the pattern of their group factor loadings will result in subgroups of clinicians internally homogeneous as to the bases they use in making clinical judgments.

Table 2

Reliability Coefficients of Undergraduate Student Judges Ranking 10 or 15 Case Histories

Number of case histories	Number of judges	Coeff. of concordance	Average interjudge correlation	Group reliability	Average corr. with psychologists	Group corr. with psychologists
10	24	.52	.51	.96	.70	.97
15	19	.32	.28	.88	.47	.76
Significance of difference (t)			2.68*		3.37*	

* Significant at the .01 level.

Naturally this is an hypothesis that was not tested by the present study, but one that is capable of being tested.

The technique of developing a factorially homogeneous set of clinical material by factor analyzing expert's judgment appears to be generalizable to the judgments of naive, less experienced judges. The student *S*s ranking the homogeneous case histories showed a higher average level of agreement among themselves and also as a group correlated higher with the judgments of the clinical psychologists. Probably student *S*s have just as many, if not more, different bases and criteria for judging clinical material as do experienced clinical psychologists and the experimental procedure of selecting clinical material that experienced judges using dissimilar judgmental criteria can agree upon also results in greater agreement among less experienced judges.

The present study was designed solely to test whether the factor analytic procedure herein described would help in developing a homogeneous set of clinical stimuli. These case histories can now be used in further research on situational variables that affect the reliability of clinical judgments quantified by the ranking method.

Summary

The study was concerned with whether inverse factor analysis procedures could be used to develop a homogeneous set of clinical case histories. Seven clinical psychologists ranked for adjustment level 15 abstracted clinical case histories of children. The matrix of clinician intercorrelations was factor analyzed and three group factors obtained which accounted for 80 per cent of interjudge variation in judgment. Factor scores were determined for each case and 10 with the smallest intracase variance in factor scores selected as a more homogeneous group. Undergraduate student *S*s ($N = 43$) then ranked for adjustment level either the original 15 or the selected 10 case histories. The student *S*s ranking the 10 cases were shown to have a significantly higher average interjudge correlation and to correlate higher with the judgments of the clinical psychologists.

Received June 20, 1955.

References

1. Bendig, A. W. Inter-judge vs. intra-judge reliability in the order-of-merit method. *Amer. J. Psychol.*, 1952, **65**, 84–88.
2. Bendig, A. W. The reliability of adjustment rating and the length of case histories. *J. consult. Psychol.*, 1955, **19**, 463–467.
3. Bendig, A. W., & Sprague, J. L. Rater experience and the reliability of case history ratings of adjustment. *J. consult. Psychol.*, 1954, **18**, 207–211.
4. Cummings, S. T. The clinician as judge: the reliability and validity of judgments inferred from Rorschach single-card performance. *J. consult. Psychol.*, 1954, **18**, 243–247.
5. Edwards, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950.
6. Hamlin, R. M. The clinician as judge: implications of a series of studies. *J. consult. Psychol.*, 1954, **18**, 233–238.
7. Hartley, H. O. Approximate tests for comparisons of rank correlations. Paper read at Amer. Statist. Ass., Washington, December, 1953.
8. Kendall, M. G. *Rank correlation methods.* London: Griffin, 1948.
9. Lyerly, S. B. The average Spearman rank correlation coefficient. *Psychometrika*, 1952, **17**, 421–428.
10. Whitmyre, J. W. The significance of artistic excellence in the judgment of adjustment inferred from human figure drawings. *J. consult. Psychol.*, 1953, **17**, 421–424.

Journal of Consulting Psychology
Vol. 20, No. 1, 1956

The Relationship Between Measures of Individual and Mass Empathy

Ralph D. Norman and Waldemar C. Leiding
The University of New Mexico

An exceedingly important problem in clinical psychology and in all interpersonal relations is the understanding of other persons, a problem on which there recently has been a rapidly growing literature, summarized quite ably by Taft (6). Among the various methods reviewed by him to measure the ability to judge people has been that of prediction of behavior or life-history data. As Taft describes this method, the judge has some acquaintance with or is given data about the subject, and he is required to predict the subject's performance on various test items or his responses to attitude and personality tests, or to predict certain aspects of the subject's life history. These are the tests which have been labeled "empathy" tests.

Quite a number of these instruments have been developed, as Taft indicates. One of the pioneer measures in this field is that of Dymond (1) whose test is made up of 4 parts, each containing the same 6 items. In the first part, the individual is asked to rate himself on a 5-point scale on each of 6 characteristics. In the second part he is asked to rate some other individual on the same 6 traits. In the third, he is required to rate the other individual as he believes this other would rate himself. In the fourth, he must rate himself as he thinks the other would rate him. In other words, according to Dymond, if two individuals, A and B, were being tested for their empathy with each other, the procedure would be as follows:

A. Part 1. A rates himself (A).
  2. A rates B as he (A) sees him.
  3. A rates B as he thinks B would rate himself.
  4. A rates himself (A) as he thinks B would rate him.

B. Part 1. B rates himself (B).
  2. B rates A as he (B) sees him.
  3. B rates A as he thinks A would rate himself.
  4. B rates himself (B) as he thinks A would rate him.

Therefore, a measure of A's empathic ability is derived by calculating how closely his predictions of B's rating (A3 and A4) correspond with B's actual ratings (B1 and B2). Similarly, a measure of B's empathy with A is obtained by calculating how closely his predictions of A's ratings (B3 and B4) correspond to A's actual ratings (A1 and A2). Since each rating is made on a 5-point scale, the test is scored in terms of the total number of points the individual is in error in his predictions. This is called the deviation score. The 6 traits used as items in all 4 parts of the test are self-confidence; superior-inferior; selfish-unselfish; friendly-unfriendly; leader-follower; sense of humor.

The Dymond test may be termed a measure of *individual* empathic ability, in that the judge is required to make predictions concerning another *individual*. In other words, according to the test author, he is required to transpose himself into the thinking, feeling, and acting of another person, to structure the world as he does, or to emphathize with him as an *individual*.

Now, another approach to the ability to judge other people may be termed that of "mass empathy." In this approach, the judge is required to predict the combined response of a group of individuals, rather than those of a single person. Taft (6) cites at least 10 studies in which this method was used, but does not include that of Norman and Ainsworth (5), owing to the comparative recency of the latter study. These investigators developed a procedure for the measurement of mass empathy which involved administration of a personality inventory (the Guilford-Martin Inventory of Factors GAMIN) in two forms. The first form was given in the usual way with the subject answering for himself. (An exception to the regular administration was the elimination of the "don't know" alternative, so that subjects were forced to answer "Yes" or "No.") Two weeks following the presentation of the first test, a second form was given to the subjects. The same questions were used, except that instead of answering for themselves, the subjects were required to answer the questions as they thought most other people of their own age and sex would answer them. For example, question 7 on the original test reads, "Do you express such emotions as delight, sorrow, anger, and the like readily?" On the second form this question was changed to read, "Do you think most people of your own age and sex express such emotions as delight, sorrow, anger, and the like readily?" Because it was felt there would be more uncertainty in answering questions for others, the question or "don't know" alternative was omitted from both forms.

The mass empathy score is derived as follows. If 51 per cent or more of the group answered in a given direction (e.g., "Yes") on a particular question on the first test (i.e., when taking it for themselves) and a particular subject answered in the same direction on the second test (i.e., when he was required to judge about most other people), he was given a point for empathy. In other words, he was "empathic" because he had made an accurate statement about what the majority (51%) had said about themselves. A subject's mass empathy score is simply the total number of points so earned.

However, in both these methods of measuring empathy, there is the possibility of some error. Hastorf and Bender (3) have pointed out that in attempting to predict the verbal responses of another person on a rating scale or personality test, there is the spurious effect of projection.[1] They ask whether or not the prediction being made is closer to the responses of the person predicted for (empathy), or to the predictor's own score (projection). They therefore attempted to isolate the factor of projection from that of empathy in a study of their own, a process accomplished by taking the difference between the empathy and the projection scores. From their results, they emphasize the fact that part of the successful prediction of another person's responses may be due to projection rather than empathy, and that a refined measure of empathic ability might approximate more adequately the psychological aspects of empathy. In the present study, an attempt was made to account for this factor of projection, as will be indicated later on.

The basic problem investigated in the present study was to explore the relationships between the Dymond test of individual empathy and the Norman-Ainsworth approach to measurement of mass empathy, using both raw and refined measures.

Subjects and Procedure

Subjects used for this study were 47 members of a social fraternity at The University of New Mexico. These men were divided into 5 groups. Three groups, consisting of 11, 11, and 10 individuals, respectively, consisted of active members of 1 year membership or more; and 2 groups, of 9 and 6 individuals each, were composed of newly initiated members. Selection in all groups was made by random choice. Each member was instructed to rate every other member of the group to which he belonged on all scales of the Dymond test. Since groups of different sizes were used, an average empathy score was obtained for each subject.

For the mass empathy testing, the same 47 subjects were given the GAMIN twice according to the procedure described above. In order that greater confidence might be placed in the "51% criterion" obtained from the use of the GAMIN, the data of Norman and Ainsworth (5) in their previous study were incorporated with the data of the present study. Since their N was 74, and ours was 47, the increased N of 121 on which the percentages were based was obviously more satisfactory. It was found that this procedure was justified, since out of 186 items on the GAMIN, there were 174 or 93.5 per cent which yielded no significant differences between the two studies.

---

[1] The use of the term "projection" by Hastorf and Bender (3) differs from that of Norman and Ainsworth (5) in their study. The former refer to it as a simple attribution of one's own traits to others. The latter defined it operationally in its more classical psychoanalytical sense, namely, denial of undesirable traits in oneself and attribution of them to others, or defense projection. In this study, we adhered to Bender and Hastorf's usage.

As mentioned above, the factor of simple projection probably enters into the operational measures of empathy used in this study according to the reasoning of Hastorf and Bender (3). Following their suggestion, it was decided to attempt to isolate projection on both tests. Thus we secured 3 measures—raw empathy, projection, and refined empathy. In obtaining the refined measure of empathy for the Dymond test, the following method was used: For each individual rated, the difference was taken between the score obtained on Part 3 of the test (wherein A rated B as he thought B would rate himself) and the score on Part 2 (wherein A rated B as he, A, saw B). Similarly, the score obtained on Part 4 (wherein A rated himself as he thought B would rate him) was subtracted from the score obtained on Part 1 (wherein A rated himself). The total projection score for all subjects was subtracted from the total raw empathy score, and this difference in score, divided of course by the number of individuals rated, gave an average refined individual empathy score for each subject. The refined mass empathy score was obtained in a similar fashion, i.e., by subtracting one point from the raw empathy score every time an individual answered an item about "most other people" the same way he answered the item about himself.

Results

The different variables explored in the present study were intercorrelated, with results as given in Table 1. The latter part of Table 1, containing the correlations between the two approaches to the measurement of empathy, indicates conclusively that there is no relationship between them. The coefficients are all in the vicinity of zero and do not differ significantly from that figure. However, the Dymond Individual Empathy Test reveals a high relationship between raw and refined empathy, with the factor of projection entering somewhat into the raw empathy score. It is probably advisable, though, that the projection variable be eliminated with future use of this test, since the correlation is significantly negative (– .47) between the refined empathy score and the projection score. With the Mass Empathy Test, on the other hand, the projection variable is a serious one with which to contend, since the relationship between it and raw empathy is very high (r of .86); the correlation between it and refined empathy is also high, the r being – .69. The resulting relationship between raw and refined empathy tends to be a negative .21, although this is not statistically significant.

Table 1

Correlations Between Various Measures of Empathy and Projection

	r
Dymond Individual Empathy Test	
Raw and Refined Empathy	.81**
Raw Empathy and Projection	.13
Refined Empathy and Projection	–.47**
Mass Empathy Test	
Raw and Refined Empathy	–.21
Raw Empathy and Projection	.86**
Refined Empathy and Projection	–.69**
Individual vs. Mass Empathy Tests	
Raw and Raw Empathy	.07
Refined and Refined Empathy	–.06
Projection and Projection	–.10

** Significant at 1% level.

Discussion

The findings of the present study regarding the lack of relationship between the measures of individual and mass empathy are in agreement with results of Hall and Bell (2) who inquired into the agreement between the Dymond test and the Kerr Empathy Test (4). The latter is also a purported measure of mass empathy, requiring the subject to (*a*) rank 14 common types of music in the order of their popularity among office workers; (*b*) rank 15 magazines in order from most to least paid circulation; and (*c*) take the role of an individual 40 years of age and rank 10 experiences from most to least annoying. Hall and Bell feel that their test aims to measure the individual's ability to assume the hypothetical average person's structural field, i.e., somehow to combine a series of "others'" individual fields into an average. (The mass empathy approach used here seems to do very much the same when the subject is asked to make a judgment about "most other people.") Hall and Bell found an average correlation of .02 between the Dymond and Kerr tests. This compares with the very low *r*'s found in the present study between the two measures of the empathic process.

Regarding the reason for the low correlation between the mass approach and the individual approach, Taft (6) has made the distinction between what he terms (after Wallin) analytic and nonanalytic judgments. He says, "In analytic judgments, the judge (J) is required to conceptualize, and often to quantify, specific characteristics of the subject (*S*) in terms of a given frame of reference. This mainly involves the process of inference, typical performances of J being traits, writing personality descriptions, and predicting the percentage of a group making a given response. In nonanalytic judgments, J responds in a global fashion, as in matching persons with personality descriptions and in making predictions of behavior. An empathetic process is usually involved in nonanalytic judgments" (6, p. 1). Later, he says, "It is suggested that the 'mass-empathy' test of prediction is more likely to be tackled analytically than is the empathy test, as it does not lend itself so readily to empathizing with any particular person. Thus the empathy test will be regarded as nonanalytic, and the mass-empathy test as analytic . . ." (6, p. 3).

Taft's distinction between the analytic and nonanalytic modes of response suggests further explanation for some of the correlations found in the present study. In the Dymond test, the rater or judge is forced to think about a particular individual with regard to a particular trait, especially in our study wherein all subjects knew each other quite well. In this situation, the factor of personal projection probably enters less and so we have the very high correlation between raw and refined empathy and the slight correlation between raw empathy and projection. On the other hand, in the mass approach, wherein the subject must make judgments about "most other people," involving, as Taft says, "a process of inference," the subject, without a frame of reference of a particular person, is quite likely to project his own feelings and attitudes into his answers. Hence the very poor *r* between raw and refined empathy and the very high *r* of .86 between raw empathy and projection.

In any case, Hastorf and Bender's criticism (3) is well taken, for the projection factor seems to enter significantly in both tests, to a much greater degree, to be sure, in the mass approach than in the individual approach.

Summary and Conclusions

Using 47 members of a social fraternity as subjects, two measures of empathy were correlated. These were the Dymond Test, a measure of individual empathy, and a procedure devised by Norman and Ainsworth for the measurement of mass empathy. In view of general criticism of these approaches in that they did not isolate the factor of projection, there were three scores derived from both measures—raw empathy, projection, and refined empathy. Correlations were about zero between the individual and mass empathy measures for all these variables, but significant correlations were obtained between these variables within the measures. Projection does enter as a factor in both tests, more seriously in the mass approach than in the individual approach, probably because there are more inferential judgments in the former than in the latter.

Received June 1, 1955.

References

1. Dymond, Rosalind F. A scale for the measurement of empathic ability. *J. consult. Psychol.*, 1949, **13**, 127–133.
2. Hall, H. S., & Bell, G. B. The relation between two tests of empathy: Dymond's and Kerr's. *Amer. Psychologist*, 1953, **8**, 361–362. (Abstract)
3. Hastorf, A. H., & Bender, I. E. A caution respecting the measurement of empathic ability. *J. abnorm. soc. Psychol.*, 1952, **47**, 574–576.
4. Kerr, W., & Speroff, B. *The empathy test.* Chicago: Psychometric Affiliates, 1951.
5. Norman, R. D., & Ainsworth, Patricia. The relationships among projection, empathy, reality, and adjustment, operationally defined. *J. consult. Psychol.*, 1954, **18**, 53–58.
6. Taft, R. The ability to judge people. *Psychol. Bull.*, 1955, **52**, 1–23.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

A Reanalysis of Beck's "Six Schizophrenias"[1]

John J. Conger, William L. Sawrey
University of Colorado School of Medicine

and Leonard F. Krause
University of Denver

The variations in symptomatology, previous history, and ultimate prognosis of patients diagnosed as schizophrenic have led a number of investigators to question whether the term itself has any fixed meaning. Fenichel, for example, remarks "Occasionally it has been doubted whether a comprehensive orientation is possible at all and whether the diverse schizophrenic phenomena actually have any thing in common. The label 'schizophrenia' is applied to so many things that it is not even of value for the purpose of prognosis. . . . Certainly, 'schizophrenia' is not a definite nosological entity, but rather embraces a whole group of diseases" (2, p. 415).

In order to lend schizophrenic diagnoses more precise connotations in terms of etiology, symptomatology, and prognosis, various investigators have attempted to divide the overall diagnosis of schizophrenia into subcategories. The number of differentiated subtypes with which they have emerged have ranged from the "process"–"reactive" dichotomy urged by some investigators, to the traditional four subtypes of simple, hebephrenic, catatonic, and paranoid. None of these classifications, based in large measure on symptomatology, has proved highly effective in practice, either in terms of prediction or control (1).

Recently, dissatisfaction with the usefulness of existing approaches led Beck and his associates to conduct an interdisciplinary study aimed at differentiating schizophrenic cases on an empirical basis—principally, as he says, in terms of variations in "ego functioning." A preliminary outline of possible variations in psychological defenses and ego functions was prepared by a number of psychiatrists under Grinker. This outline was then broken down by Stephenson into 120 separate items descriptive of personality functioning. Q sorts of these items were then made by psychologists and psychiatrists for groups of outpatients previously diagnosed as schizophrenic. The psychiatrists (referred to as P in this study) sorted the items on the basis of clinical interviews with the patients, the psychologists (referred to as A and B) on the basis of individual Rorschach protocols. Each patient was rated by one psychiatrist and at least one, usually two, psychologists. All of the sorts for a group of patients were then intercorrelated and a factor analysis made of the resulting matrix.

Three such analyses were made. Initially Beck and his co-workers began by rating 20 patients, 12 adults and 8 children. Since an inordinately large correlation matrix would have resulted had an attempt been made to factor-analyze the total group, it was subdivided "into an 'agreement' group, i.e., one in which psychiatrists and Rorschach test investigators showed agreement in judging the patients; and a 'disagreement' group" (1, p. 71). Apparently, factor analyses of the intercorrelations between raters with respect to individual patients were used as the basis for assignment to the agreement or disagreement group. Presumably when all the raters had similar factor loadings on a patient, he was assigned to the agreement group. When they did not, he was assigned to the disagreement group. The exact degree of similarity required for assignment to the agreement group could not be ascertained from the information available. However, real questions arise in this connection, since, as Shaffer points out, ". . . the psychiatrists and psychologists seemed in better agreement on 3 of the 8 'disagreement' cases than on 3 of the 12 'agreement' ones" (3, p. 472). At any rate, the division was made primarily for purposes of convenience in handling the data. In addition to these two groups, a third group of 20 schizophrenic children, some of them repeaters from the first two groups, was also studied.

---

[1] Based on a paper presented at the annual meeting of the Midwestern Psychological Association, Chicago, Illinois, 1955.

On the basis of these three analyses, Beck and his collaborators derived five factors. These factors, then, led them to define, "by their juxtaposition," six types or classes of schizophrenic cases (1, p. *226*).

The Present Problem

While we are not concerned here with the merits of Q technique, we question the methodology employed in obtaining Beck's results, and feel that it casts serious doubts on the validity of his conclusions. In the case of each of the above analyses, the group correlation matrix consisted of X individual ratings of Y patients. Thus, each correlation coefficient represents a combination of rater and patient characteristics. In other words, the amount of variance the correlation accounts for is distributed between the two in indeterminate amounts. Therefore, when this correlation matrix is factor-analyzed it is virtually impossible to tell to what extent the factors represent raters or patients.

Procedure

Another method of analysis might separate the effects of raters and patients. Assume that from Beck's over-all matrix we extract the individual intercorrelation matrices for the three raters, and factor them separately. If Beck's factors are in fact due primarily to patient differences, we should then emerge with three quite similar factor patterns. This was done, in each case attempting to rotate toward simple structure in a fashion as similar to Beck's as possible.

Table 1

Rotated Factor Matrix, Rater A

Subject number†	Factor I	Factor II	Factor III
3		.82	
5		.77	
13		.49	.41
14		.50	.44
8	.74		
19	.74		
4	.41	.54	

Note.—Loadings with an absolute value under .40 are omitted.

† These numbers correspond to the numbers used by Beck.

Table 2

Rotated Factor Matrix, Rater B

Subject number†	Factor I	Factor II	Factor III
3		.86	
5		.76	
13		.60	.69
14		.45	.62
1	.87		
9	.83		
11	.80		
12	.87		
15	.65		
8	.88		
19	.77		
4	.61		−.41

Note.—Loadings with an absolute value under .40 are omitted.

† These numbers correspond to the numbers used by Beck.

Results and Interpretation

Agreement group. From his analysis of the over-all agreement group, Beck concluded that there emerged four factors, defining three schizophrenic types, S-1, S-2, and S-3. A somewhat different picture emerged when we analyzed each of the three raters' matrices separately. Rater A's rotated factor matrix is shown in Table 1. Assuming an absolute factor loading of .40 or larger to be meaningful enough to help identify a factor, it is clear that individuals 8, 19, and 4 identify Factor I. Similarly, Factor II is identified by individuals 3, 5, 13, 14, and 4, and Factor III by individuals 13 and 14. After extraction of these three factors, it can be seen that none of the remaining loadings are great enough to account for further factors.

The corresponding matrix for Rater B is shown in Table 2. In this case, Factor I is identified by individuals 1, 9, 11, 12, 15, 8, 19, and 4. Factor II is identified by individuals 3, 5, 13, and 14. The only remaining factor, Factor III, is identified by individuals 13, 14, and 4.

Rater P's matrix is shown in Table 3. Here Factor I is identified by individuals 1, 9, 11,

Table 3

Rotated Factor Matrix, Rater P

Subject number†	Factor I	Factor II	Factor III	Factor IV
3		.74		
5		.84		−.42
13		.65	.45	
14			.59	
1	.69			−.63
9	.64			
11	.61			
12	.68			
15	.63	−.53		
8	.72			
19		.51		
4	.44			

Note.—Loadings with an absolute value under .40 are omitted.

† These numbers correspond to the numbers used by Beck.

12, 15, 8, and 4. Similarly Factor II is identified by individuals 3, 5, 13, 15, and 19. Factor III is identified by individuals 13 and 14. Finally, the remaining factor, Factor IV, is identified by individuals 5 and 1.

How similar are these individual factor matrices?

It should be immediately apparent that there exists some disagreement between the various raters as to the number of factors (and consequently the possible number of types of schizophrenia) present in this population. Raters A and B, for example, emerge with only three, while Rater P finds four. Furthermore, there appear to be interrater differences in terms of the individuals identified by apparently parallel factors. The corresponding factor patterns for each of the three raters are presented in Table 4.

As may be seen, there is only fair agreement among all three raters in defining Factor II. They all assign moderate or high loadings on this factor to individuals 3, 5, and 13, and negligible loadings to individual 8. Furthermore, Raters B and P agree in assigning negligible loadings to individuals 1, 9, 11, and 12 (these individuals were not rated by Rater A); however, they disagree on individual 15. In addition, the three raters disagree on individuals 14, 19, and 4. Raters A and B, for example, assign a moderate positive loading to individual 14, while Rater P gives him a negligible loading. While Raters A and B agree in assigning a negligible loading to individual 19, Rater P gives him a moderate positive loading. In the case of individual 4, Raters B and P agree in assigning him a negligible loading, while Rater A gives him a moderate positive loading.

On Factor I, there is somewhat better agreement than on Factor II. All three raters agree that individuals 3, 5, 13, and 14 have negligible loadings on this factor and that indi-

Table 4

Combined Factor Patterns of all Raters for the "Agreement" Group

Subject number†	Factor I			Factor II			Factor III			Factor IV
	A	B	P	A	B	P	A	B	P	P
3	0	0	0	8	9	7	0	0	0	0
5	0	0	0	8	8	8	0	0	0	−4
13	0	0	0	5	6	7	4	7	5	0
14	0	0	0	5	5	0	4	6	6	0
1	*	9	7	*	0	0	*	0	0	−6
9	*	8	6	*	0	0	*	0	0	0
11	*	8	6	*	0	0	*	0	0	0
12	*	9	7	*	0	0	*	0	0	0
15	*	7	6	*	0	−5	*	0	0	0
8	7	9	7	0	0	0	0	0	0	0
19	7	8	0	0	0	5	0	0	0	0
4	4	6	4	5	0	0	0	−4	0	0

Note.—Digits in the above table indicate factor loadings expressed to a single significant figure and with the decimal points omitted. Loadings below an absolute value of .40 are indicated as 0.

† These numbers correspond to the numbers used by Beck.

* These subjects were not rated by A.

viduals 8 and 4 have moderate to high positive loadings. In addition, Raters B and P agree in assigning positive loadings to individuals 1, 9, 11, 12, and 15. In the case of individual 19, however, Raters A and B disagree strongly with Rater P. While A and B both assign him a high positive loading, P gives him a negligible loading. Whether there would have been additional agreement or disagreement if Rater A had rated all subjects, we do not know.

On Factor III, all three raters agree in assigning negligible loadings to individuals 3, 5, 8, and 19, and moderate positive loadings to individuals 13 and 14. Raters B and P agree in assigning negligible loadings to individuals 1, 9, 11, 12, and 15. However, while Raters A and P assign negligible loadings to individual 4, Rater B assigns him a moderate negative loading.

When the above three factors are removed from the various factor matrices, the situation becomes more complicated. Raters A and B have, in effect, no Factor IV; that is, on no individual do they have any remaining factor loadings above .40. Rater P has a fourth factor, identified negatively by individuals 5 and 1.

The above comparisons make it clear that there is disagreement between the various raters both as to the number of factors present in this patient population, and also as to which persons identify apparently parallel factors. In his analysis of the over-all agreement intercorrelation matrix, Beck obtained four factors. Since in our analysis of the individual rater's matrices, Rater P is the only one having a fourth factor, it appears that Beck's fourth factor was probably a result of disagreement among raters and confounding of raters and individuals.

It is also possible to view Table 4 in terms of the amount of rater agreement as to the factors characterizing each patient. For example, it can be seen that all raters agree that individual 3 is characterized strongly by Factor II, and not by Factor I or III. On the other hand, there is relatively poor rater agreement as to the factors characterizing individual 4. Raters B and P agree that individual 4 has negligible loadings on Factor II. Rater A, however, gives him a moderate positive loading on this factor. On Factor III, Raters A and P give this patient negligible loadings, while Rater B gives him a moderate negative loading. All agree that Factor I characterizes this patient. Rater P assigns individual 4 a negligible loading on his Factor IV.

Let us first consider the seven patients on whom ratings by all three judges were available. Since three factors were extracted, there are 21 possible instances in which all three raters may agree. However, in only 15 of these instances does such agreement occur, even when all positive loadings above .40 are considered identical and when anything from − .39 to + .39 is considered to be zero in the table. For the remaining five patients who were rated by only two raters, there are 15 possible instances of agreement. If we consider all positive loadings above .40 as identical for our purposes, there are 14 instances of agreement.

In our opinion, the limited agreement among raters shown in the above analyses does not justify Beck's contention that clear-cut factors, based on patient differences, emerge from his analysis of the "agreement" group.

Disagreement group. It is somewhat difficult to interpret Beck's statements as to the factors which he feels emerge from an analysis of his disagreement group. At one point, he states that four factors emerge, but three of these are the same Factors II, III, and IV

Table 5

Combined Factor Patterns of all Raters for the "Disagreement" Group

Subject number†	Factor I			Factor II			Factor III
	A	B	P	A	B	P	B
2	6	7	5	0	0	0	0
6	0	0	6	0	−7	0	0
7	5	0	0	0	0	5	−8
10	6	7	5	0	0	0	−6
16	0	0	0	−7	−7	−6	0
17	0	0	5	−7	−8	0	0
18	7	7	0	0	0	−7	0
20	7	6	0	0	0	6	−5

Note.—Digits in the above table indicate factor loadings expressed to a single significant figure and with the decimal points omitted. Loadings below an absolute value of .40 are indicated as 0.

† These numbers correspond to the numbers used by Beck.

of his agreement group. We are somewhat at a loss to determine how he arrived at this conclusion. Since he included in his disagreement rotated factor matrix three inputs labeled S-1, S-2, and S-3, it would appear that these were somehow obtained from his agreement analysis, and included as reference variables. However, if this were actually the case, these inputs should load up on different factors in the disagreement analysis. Actually, however, they all load up on Beck's disagreement Factor II. Because of these and similar problems, we do not feel capable of commenting on Beck's analysis and interpretation of the disagreement group.

However, we were again able to extract and analyze the individual rater's intercorrelation matrices as in the case of the agreement group. From the results, the factor patterns shown in Table 5 were prepared. As may be seen, Rater A emerges with two factors, Rater B with three factors, and Rater P with two factors. It seems to us that these data emphasize the distortions that may result from analyzing an over-all intercorrelation matrix which confounds raters and patients. For while no individual rater emerges with more than three factors, Beck isolated four factors in his over-all analysis. It does not seem fruitful to investigate rater agreement further, since this group was explicitly selected on the basis of disagreement. However, one again wonders about Beck's criteria for disagreement since all raters were in agreement as to the factor loadings of individuals 2 and 16.

The data from Beck's analysis of his third group of 20 schizophrenic children are not available, and consequently cannot be commented upon.

Summary

In the present analysis we have demonstrated (*a*) that there is disagreement among raters as to the number of factors present in this population of patients, and (*b*) that there is also disagreement as to which persons identify apparently parallel factors.

In light of this evidence of interrater unreliability, we do not feel that Beck is justified in his conclusions as to the factors, and consequently the schizophrenic types which characterize his population.

Received June 3, 1955.

References

1. Beck, S. J. *The six schizophrenias.* Res. Monogr. No. 6. New York: Amer. Orthopsychiat. Ass., 1954.
2. Fenichel, O. *The psychoanalytic theory of neurosis.* New York: Norton, 1945.
3. Shaffer, L. F. Review of S. J. Beck, *The six schizophrenias. J. consult. Psychol.*, 1954, **18**, 472–473.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

A Reply to the Reanalysis of Beck's "Six Schizophrenias"

William Stephenson
Greenwich, Connecticut

The paper by Conger, Sawrey, and Krause (2) reanalyzes data in Beck's monograph (1) and questions the initial analysis on the ground that it involves a combination of rater and patient variances in unknown or unknowable proportions. Their own alternative analysis results, in one case, in agreement being shown in 15 out of 21 instances, and in another case in 14 out of 15. They conclude that this is "limited" agreement, and evidence of "interrater unreliability." Having regard to the complex matter involved, the present author wishes that he could be as successful on the racecourse as Beck's colleagues were in the clinic. Other considerations suggest, too, that the sounder conclusion, even when the data are analyzed in a different way, is that considerable evidence is forthcoming for Beck's hypotheses.

The point at issue is as follows. The reanalysis brings into focus rater *specificities* rather than factors of more general range. If a matrix is factored for a particular Rater A, factors are apt to involve his specificities, due to his greater or lesser training and the like; and while it may be of interest to study such factors, they are unlikely to be of immediate interest or necessarily relevant to one's more general theoretical questions. Clearly, in this connection, the reanalysis involves itself only in certain *factors,* and not in their *factor-arrays,* without which Conger, Sawrey, and Krause are working completely in the dark. Beck, however, was working in the full daylight of his factor-arrays, and not with mere factors as such or the possibility of specificities. His concern was with the general question as to whether *different* psychologists and psychiatrists, over a range of *different* patients, could provide factors of theoretical interest—the latter matters being judged by the factor-arrays at issue. The randomizing of raters and patients involves the sort of logic with which agronomists are familiar when, in their experiments, they divide a field into plots in such a way as to level out differences of fertility, shadow, and the like in different parts of the field. In the clinical case we seek to cancel out the specificities of either raters or patients, with the firm hope that generality is the more likely that way.

The trouble, if I may say so, is that it is all too easy to slip into categorical mistakes in factor studies. To talk of "interrater unreliability" is to involve oneself in such a mistake. It would have been quite easy to provide Beck and his colleagues with a Q sample which would result in one hundred per cent agreement about patients every time, but about quite unimportant matters. To reach a factor of real psychological interest (as judged by the factor-array), under the complex conditions of Beck's Q sample, even if the loadings are relatively small and their deployment over different raters somewhat patchy, may be altogether far more important than to reach a factor with high loadings for everyone concerned. The latter factor would probably only deal with the obvious, the banal, or worse. With the Q sample used by Beck, involving as it does some complex theoretical matters, the wonder is (as far as this author is concerned) that there was *any* agreement between psychologists themselves, and with the psychiatrists, rather than that the agreement is "limited."

The same kind of categorical slip is found, as well, in the critical remarks made by these

Table 1

Intercorrelations and Factor Loadings for Two Patients

Patient 1

Rater	1	2	3	4	Factor loadings
1. Psychiatrist		.16	.16	.16	.40
2. Psychologist A			.16	.16	.40
3. Psychologist B				.16	.40
4. Psychologist C					.40

Patient 2

					Factor loadings	
Rater	1	2	3	4	I	II
1. Psychiatrist		.25	.25	.25	.50	.00
2. Psychologist A			.25	.61	.50	.60
3. Psychologist B				.25	.50	.00
4. Psychologist C					.50	.60

authors, and by Shaffer earlier (3), about the way in which Beck divided his 20 patients into an "agreement" and a "disagreement" group. These categories were determined not by *size* of correlation (a criterion that the critics appear to recommend) but by comparability of *factors*. The examples in Table 1 will make the difference clear.

The average correlation for Patient 1 is smaller than for Patient 2 but the raters are in agreement about the factor at issue; in the other patient, however, though they are apparently in agreement about one factor, they differ about the other. It is easy to see that the situation in the latter case could be such that the raters could agree about Factor I by an amount less than 0.40, with Factor II correspondingly higher for all four raters. Everything will depend upon how these Patients 1 and 2 appear in a table including other patients as well. The assumption is reasonable, however, that matters are likely to be straightforward for Patient 1, although, with other patients and raters in the matrix, a complex factor situation is still likely, whereas for Patient 2 matters are already complicated. It was with such considerations in mind, then, and not the mere size of intercorrelations, that Beck's patients were divided into two categories for convenience.

It is well known that multiple factor conditions are permissive (4). An appropriate factor methodology, therefore, is to use factor analysis to answer psychological questions, in a deductive rather than an inductive setting. In Beck's case the concern was with the question whether ratings were, or were not, consistent with a certain theoretical position. The same data might be consistent with other theoretical positions as well, but this would be a matter for separate consideration. One should not think of such factor analysis, therefore, as a producer of facts for inductive regard (which is still another categorical mistake), but as a reproducer of facts, from a deductive standpoint.

We see, then, in the short space of a single paper, several sources of categorical assumption, against which Q method is always directed. Beck's initial analysis made no such mistakes, and is clearly on sound lines.

Received October 24, 1955.
Early Publication by the Editor.

References

1. Beck, S. J. *The six schizophrenias.* Res. Monogr. No. 6. New York: Amer. Orthopsychiat. Ass., 1954.
2. Conger, J. J., Sawrey, W. L., & Krause, L. F. A reanalysis of Beck's "Six schizophrenias." *J. consult. Psychol.*, 1956, **20**, 83–87.
3. Shaffer, L. F. Review of S. J. Beck *The six schizophrenias. J. consult. Psychol.*, 1954, **18**, 472–473.
4. Stephenson, W. *The study of behavior; Q-technique and its methodology.* Chicago: Univer. of Chicago Press, 1953.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Reply to Dr. Stephenson

John J. Conger, William L. Sawrey
University of Colorado School of Medicine
and Leonard F. Krause
University of Denver

While we do not wish to belabor discussion of the methodological differences existing between Stephenson (4) and ourselves (2), we do feel that several specific points in his reply require comment. In the first place, Stephenson remarks that even with our method of reanalyzing Beck's data, we emerge, in the one case, with 15 out of 21 possible instances of rater agreement (when three raters were used) and, in another case, with 14 out of 15 possible instances of agreement (when only two raters were used). Two points should be stressed in this connection. First, agreement of this magnitude occurs only when all positive loadings above .40 are considered identical and when anything from minus .39 to plus .39 is considered to be zero, i.e., also identical. Second, this amount of agreement obtains only for Beck's "agreement" group, which was presumably preselected for interrater agreement. When the same criteria for rater agreement are applied to the two factors which all raters had in Beck's "disagreement" group, agreement occurs in only 6 out of 16 possible instances, as Table 5 of our paper (2) demonstrates.

Stephenson remarks that he wishes he could do as well at the race track as Beck and his colleagues have done in the clinic. In this connection, we would like to make it clear that we have only admiration for the clinical skills displayed by Beck and his colleagues, and for their ingenuity in attempting to conceptualize the problem of schizophrenia in dynamically more meaningful terms. Our concern here is not a clinical, but solely a methodological one, having to do specifically with the manner in which the data were analyzed. Parenthetically, however, Stephenson's race-track analogy does not impress us as particularly relevant, since in order to make visits to the race track extremely profitable one need do only slightly better than chance. A similar degree of success in the clinic does not seem to be a particularly desirable goal.

However, as Stephenson himself points out, the essential point of difference between us does not concern the extent of interrater agreement in Beck's data. It concerns, rather, the theoretical necessity for such agreement. In Stephenson's words, "To talk of interrater unreliability" in discussing factor studies is to involve one's self in a "categorical mistake." We do not agree. As we think our reanalysis demonstrates, replicable findings from factor analysis cannot be assured unless a reasonable degree of interrater reliability exists at the outset. Stephenson argues that "if a matrix is factored for a particular Rater A, factors are apt to involve his specificities, due to his greater or lesser training and the like. . . ." This may well be, and if so, generalizing from this rater's findings would certainly be misleading. That is why it is desirable to determine whether substantial agreement among a variety of raters can be obtained before making such generalizations.

But the problem of generality cannot, in our view, be solved simply by combining the ratings of all raters across all patients, regardless of degree of agreement, into one intercorrelation matrix, and then factoring it. Stephenson argues that from such a procedure one can derive factors of more "theoretical interest." We are not entirely certain what Stephenson means by more theoretical interest, but we are sure that one cannot be confident, with this procedure, of deriving factors which, taken together in "factor-arrays," define "six types of schizophrenia." To take an extreme example, it would be possible to factor-analyze the intercorrelation matrix of the ratings of one patient by twenty raters. One might well obtain a number of factors from which factor arrays could be constructed. But we do not see how these factor arrays could be considered to define types of schizophrenia, though they might help to define types of raters.

Of course, in their analysis, Beck and Stephenson used a number of patients as well as a number of raters, but it seems to us that the argument still applies. To the extent that the raters disagree, the factors obtained will still confound raters and patients. Where there is a considerable degree of interrater disagreement (as was the case in their disagreement group), might not new raters differ from the present ones even in rating the same patients? If so, factoring their ratings might lead to different factors, and consequently to new sets of factor arrays, or "types of schizophrenia." This, however, would be manifestly unmeaningful, it seems to us, since both sets of schizophrenic types would have been derived from the same patients.

Stephenson mentions several other examples to indicate that we "are working in the dark." He points out that our reanalysis "involves itself only in certain factors, and not in their factor arrays." We fail to understand the dire implications of this for the degree of illumination in which we are working. We explicitly recognized that Beck and Stephenson were working with "factor arrays," though we referred to them as "factor patterns." But since it is the factors themselves which led these investigators to define "by their juxtaposition" the factor arrays or types of schizophrenia (1, p. 226), it seems to us only logical that changes in the factors would be reflected in changes in the arrays.

In connection with Stephenson's reference to "the sort of logic with which agronomists are familiar," we would like only to note that agronomists in their work are aware not only of the desirability of randomizing, with which we also of course agree, but of the difficulties in interpretation which may result from partial or complete confounding. This, we feel, is the primary problem in the present analysis.

We are also charged with a "categorical slip," as is Shaffer (3), in regard to the way in which Beck divided his patients into an "agreement" and a "disagreement" group. We are accused of failing to understand that these categories "were determined not by *size* of correlation . . . but by comparability of *factors*."

We do not see how Stephenson could have arrived at this conclusion when we specifically stated, "Apparently, factor analyses of the intercorrelations between raters with respect to individual patients were used as a basis for assignment to the agreement or disagreement groups. Presumably when all the raters had similar factor loadings on a patient, he was assigned to the agreement group. When they did not he was assigned to the disagreement group" (1, pp. 2–3). We further stated that the exact degree of similarity required for assignment to the agreement group could not be ascertained from the information available.

We still feel this way, especially since in some instances in our analysis, the factor loadings were more comparable for the disagreement than for the agreement group. Consider, for example, the factor loadings of Individual 19 in Table 4 of our paper, and Individual 2 in Table 5. As these tables show, the loadings were less comparable for Individual 19, a member of the agreement group,

Table 1

Intercorrelations of Raters for One Patient of the Agreement Group and One Patient of the Disagreement Group

Patient and rater	Intercorrelations 1	2	3
Patient 19			
1. Psychiatrist		.01	.20
2. Psychologist A			.42
3. Psychologist B			
Patient 2			
1. Psychiatrist		.24	.32
2. Psychologist A			.58
3. Psychologist B			

than for Individual 2, a member of the disagreement group. This does not seem surprising in view of the obtained intercorrelation matrices of the three raters for these two patients shown in Table 1 of the present paper (data concerning Psychologist C, the fourth rater, were not included in Beck's monograph and consequently cannot be included below). Although we are still uncertain as to the exact criteria for assignment to the agreement and disagreement groups, our more fundamental question concerns the usefulness of a method which fails to take into account correlation size, resulting in categorizations such as the above.

We would like to raise one final point. One of the reasons we did not proceed to analyze Beck's disagreement group to the same extent as the agreement group was that we could not understand his statement that in the disagreement group four factors emerged, but three of these are the same as Factors II, III, and IV of his agreement group. We could not see how this conclusion was reached. Since three inputs labelled S-1, S-2, and S-3 were included in the disagreement-rotated factor matrix, presumably as reference variables, it appeared to us that these were somehow obtained from the agreement analysis. If this were actually the case, however, these inputs should have loaded up on different factors in the disagreement analysis. But since they all loaded up on Beck's disagreement Factor II, we still do not understand, and Stephenson has not explained, the meaning of these inputs, or their relation, if any, to the agreement group factors.

To summarize, despite other areas of disagreement between Stephenson and ourselves, our basic and apparently irreconcilable points of difference seem to involve differing notions as to the theoretical importance of interrater reliability and confounding in a study such as the present one. In Stephenson's opinion, "to talk of interrater unreliability" is to involve one's self in a "categorical mistake." We simply cannot agree. In our opinion, it is a mistake to derive types of schizophrenia by factoring a matrix in which raters who disagree substantially with one another are confounded with patients.

Received November 15, 1955.
Early Publication by the Editor.

References

1. Beck, S. J. *The six schizophrenias.* Research Monograph No. 6. New York: American Orthopsychiatric Association, 1954.
2. Conger, J. J., Sawrey, W. L., & Krause, L. F. A reanalysis of Beck's "six schizophrenias." *J. consult. Psychol.*, 1956, **20**, 83–87.
3. Shaffer, L. F. Review of S. J. Beck, *The six schizophrenias. J. consult. Psychol.*, 1954, 18, 472–473.
4. Stephenson, W. A reply to the re-analysis of Beck's "six schizophrenias." *J. consult. Psychol.*, 1956, **20**, 88–89.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

The Effects of Perceptual Training on the Rorschach *W* and *Z* Scores[1]

Howard Leventhal
University of North Carolina

Although the Rorschach is now the most widely used instrument in American clinical psychology, it has little in the way of empirically established foundations. Though a number of Rorschach workers have recognized the need for normative data, there still remains the task of demonstrating the validity or meaningfulness of the hypotheses concerning the variables being standardized. Direct experimentation with specific factors of the test seems to have been neglected because experimentation, being analytical, is counter to the gestalt viewpoint held by many Rorschach workers. Bergmann's discussion of gestalt principles (4, pp. 450–451) makes clear, however, that the interrelationship of the Rorschach variables necessary to describing a unitary personality does not preclude experimentation with any one or any combination of variables to discover the laws of these subsystems.

The significance of the *W* response has been defined largely in terms of intellective characteristics. Rorschach (18, p. 59) wrote that "The number of *W* responses is to be considered primarily an indicator of the energy of associative activity, dependent on a dispositional set of the subject." Beck stated that (3, p. 10) "The Rorschach test projects degree or height of intelligence in two factors-whole percepts (*W*) and organization (*Z*)." The *Z* score used by Beck (3, pp. 12–13) is considered "a more accurate representative of intelligence functioning per se," as it considers intermediate values of organization missed by *W*. He feels that the *Z* scores "vary directly as the intelligence of *S*."

Previous Studies with *W* and *Z*

If *W* and *Z* scores can be meaningfully interpreted as measures of intelligence, it should be possible to demonstrate both their validity and reliability. Any valid measure of intelligence should be reliable. Thus, *W* and *Z*, as indices of intelligence, should not be unduly influenced by extraneous events like small amounts of training or slight emotional and attitudinal shifts. Since consistent results have not been obtained from correlational studies investigating the relationship of Rorschach *W* and *Z* to intelligence, experimental research directed toward locating possible variables influencing the variability of *W* and *Z* scores might provide clues as to why these scores do not correlate more highly with intelligence. Among the factors that influence *W* and *Z* responses are: Stress has been found to reduce *W* significantly (7). Examiner differences and retesting produce significant changes in *W*% (15). *W* responses increase in a social situation (14). As some set seems to be of importance, a number of experiments have attempted to study its influence by inducing various sets in subjects and measuring differences in responses. These sets have been manipulated by asking subjects to give "good" and "bad" records on different occasions. No changes in the record in *W*, and high intertest correlations, are reported in two instances (9, 10), while low intertest correlations are reported in another (5). Instructions to pay attention to certain areas of the blot (12) produced significant changes in *W*, as did hearing at a lecture that intelligent people see *W* (1). Even looking at colored advertisements before taking the Rorschach increased the number of responses (17). Keyes (13), in a more direct attack on this problem, treated four groups as follows: *1*, training on Street Gestalt figures plus instructions to see only *W*'s on the Rorschach; *2*, training on the Street Gestalt test plus standard Rorschach instructions; *3*, no training but instructions to see *W*'s only; *4*, no training but standard Rorschach instructions. Group *1* produced significantly more *W* responses than any other group while neither training alone nor instructions to give *W* only produced a significant increase in frequency of *W*'s.

---

[1] This research was completed as a master's thesis at the University of North Carolina. Thanks are extended to Drs. W. Grant Dahlstrom, June E. Chance, and George S. Welsh for their invaluable advice and criticism.

Nature of Rorschach Perceptual Processes

Considering the Rorschach as a perceptual task, we might theorize about the nature of perceptual processes involved in making *W* responses, and variables that influence these processes. Perceiving a *W* response appears basically to be a problem of organizing the blot, while perception of subareas seems to involve ignoring the total gestalt and analyzing the card into new units. These two organizing processes seem similar to perceptual factors found in Thurstone's "Factorial Study of Perception" (20). Thurstone's first factor, the *A* factor, seems to "represent the ability to form a perceptual closure against some distraction" (20, p. 101). The Street Gestalt test and the Gottschaldt figures have high loadings on this factor. The latter test, however, also has a high loading on the *E* factor, which Thurstone considers to represent "The ability to shake off one set in order to take a new one . . . it implies flexibility in manipulating several more or less irrelevant or conflicting gestalts" (20, p. 111). In contrast, the Street Gestalt Test does not have a high loading on the *E* factor.

Problem

The purpose of this study is to examine experimentally whether the Thurstone *A* and *E* factors are related to *W* and *Z* responses to the Rorschach. This relationship can be demonstrated by contrasting Rorschach responses of groups pretrained on Closure figures (factor *A*) with those trained on Gottschaldt figures (factor *E*). Responses of both trained groups will be compared to responses made by a control group receiving no pretraining.

Hypotheses

1. Training on a perceptual task before taking the Rorschach test will alter the *W* and *Z* responses of trained groups.
2. The group trained on the incomplete figures test (Thurstone's *A* factor) will produce more *W* responses and receive higher organization scores than a control group who did not receive perceptual training.
3. The group trained on the Gottschaldt test (Thurstone's *E* factor) will give fewer *W* responses and have a lower organization score than the control group.

Procedure

Forty-nine volunteer college males were used as subjects. The seating plan used during the Rorschach and the training periods was that of Harrower and Steiner (11, p.

Table 1

Experimental Design and Hypotheses

Group	*N*	Training	Rest	Rorschach performance
I	17	Closure practice using incomplete figures	5′	More *W* responses and higher *Z* scores than Groups II and III
II	13	Practice in breaking down Gestalts and forming new units on Gottschaldt figures.	5′	Fewer *W* responses and lower *Z* scores than Groups I and III
III	19	No perceptual training, verbal task	5′	*W* and *Z* scores intermediate to those of Groups I and II

119). A 300 watt Viewlex projector placed about 21 feet from the screen gave a screen area of 60 by 40 inches for Card I. Subjects were told, previous to the experiment, that they were volunteering for two experiments, one conducted by a graduate student, and one by a faculty member, and that a maximum of one and one-half hours would be needed to complete the work. They were assigned to one of three groups depending on when they were free. After completion of the training period, subjects were reminded that they were also to participate in a second experiment and told that in the meantime they might take a 5-minute rest. These precautions were taken to prevent subjects from assuming that the instructions for the pretraining period were supposed to hold for the Rorschach. Different experimenters conducted the two sessions.

Group I. Group I was trained on closure figures before being given the Rorschach. They were shown 35, double-frame 35 mm. slides of closure figures from the Mooney Closure Test (16). Figures containing any irrelevant details were excluded, and for those with white figures on a black ground, the photographic negatives were used in the slide. The figures were projected for 20 seconds each, and subjects recorded their answers in a booklet. After the first presentation, the slides were again projected for 33 seconds each. This time the experimenter gave the correct answers and helped subjects to see the figures. He emphasized that all segments must be used in order to see the figures. A 5-minute recess followed.

Group II. Group II was trained on the Thurstone Concealed Figures Test (21), a multiple-choice modification of the Gottschaldt figures. Figures were projected with an opaque projector. A figure was projected with four other more complex figures in which it might be hidden. Subjects were instructed to check all complex figures the smaller figure was hidden in. Two practice slides and 21 test items were used. All slides, except the first, were seen for 30 seconds. After the test, the items were again projected for 50 seconds each and the correct answers pointed out. Subjects were told that to see the enclosed figures, they must ignore the unity of the enclosing figure and search out the smaller figure.

Group III. The control group was given the Shipley–Institute of Living Scale for Measuring Intellectual Impairment (19). Time limits of 10 minutes each were allowed for the vocabulary and abstractions subtests.

Rorschach administration. The Rorschach was administered to all groups by projecting 35 mm. double frame Kodachrome slides of the blots. The projections were faithful reproductions of the blots. Each of the ten slides was presented for one and one-half minutes. Subjects were given answer booklets and asked to write down what they saw or anything that the slides might represent to them. During the inquiry, the subjects were given sheets of achromatic reproductions of the blots on which to encircle their responses. Examples of whole and detail responses were given; the slides were projected again during the inquiry.

Bureau Ednl. Psy. Research
COLLEGE
Acc No ....

Results

The first response made to each card was scored for *W* and for *Z*, following Beck's (2) procedure. As some subjects did not respond to all of the cards, separate analyses were also made of *W* and *Z* for the subjects who gave more than ten responses to the Rorschach.

All analyses of variance for *W* and *Z*, except for number of rejections, were statistically significant. Tests for homogeneity of variance showed nonsignificant differences (Table 2). As the direction of differences was predicted, one-tailed *t* tests were used to compute the significance of differences between means.

When compared with the Gottschaldt-trained group, the closure-trained group gave significantly more *W* and *Z* responses. This was so when *t* tests were computed using all subjects, or when subjects were omitted who gave less than 10 responses. The Gottschaldt-trained group also produced significantly fewer *W* and *Z* responses than did the control group. As predicted, the closure-trained group gave more *W* and *Z* responses than the control

Table 2

Means, Sigmas, and Significance of Variance Ratio (*F*) for *W* and *Z*

Measure	Group I	II	III	*F*	*p*
All subjects					
Mean *W*	6.21	4.31	5.65	3.45	.05
Sigma *W*	2.07	1.75	2.18		
Mean *Z*	27.79	19.23	25.68	4.56	.05
Sigma *Z*	8.00	7.11	8.66		
M Rej.	.63	.46	.53	$S_w^2 > S_b^2$	
Subjects giving more than 10 responses					
Mean *W*	6.47	4.31	5.62	3.85	.05
Sigma *W*	2.10	1.75	2.25		
Mean *Z*	29.70	19.23	25.84	6.15	.01
Sigma *Z*	7.46	7.11	8.91		

Table 3

Significance of Differences Between the Group Means for *W* and *Z*

Groups compared	Diff. *W*	*t*	*p**	Diff. *Z*	*t*	*p**	*df*
All subjects							
I–II	1.90	2.60	.01	8.56	2.96	.005	30
I–III	.56	.83	.25	2.11	.79	.25	34
II–III	1.34	1.79	.05	6.44	2.18	.025	28
Subjects giving more than 10 responses							
I–II	2.16	2.76	.01	10.47	3.48	.005	26
I–III	.84	1.14	.15	3.86	1.35	.10	29
II–III	1.32	1.71	.05	6.61	2.23	.025	27

* *p* less than value indicated with one-tailed test.

group, but the differences were insignificant (Table 3).

Discussion

The random design of this study has certain definite advantages over a test-retest design where groups are equated by means of their pretest performance. If subjects had had experience with the Rorschach, or with any other inkblot test, the response set established during this previous test might act to reduce or to eliminate the effectiveness of the set established in the training period. Since *W* and *Z* are dependent upon *R*, data analysis was confined to first responses to each slide to control for number of responses. Other possible methods of controlling *R*, matching on *R* after training and use of covariance technique (8), did not seem appropriate as it was not known how the experimental treatments might effect responsiveness, and the regression of *W* and *Z* on *R* could not a priori be expected to be linear. Asking subjects for an arbitrary number of responses was decided against, and Cronbach's (6, p. 426) suggestion of scoring a fixed number of responses was followed. If training results in changes in organization value and in number of *W* responses given, then it is reasonable to expect that the effects of such training would be as apparent on the first response to a card as on later responses. If training effects did not appear in first responses, or if they were not demonstrable with the fairly small number of subjects used, then they would probably be of little practical importance.

Exactly what was transferred from training to test period is unanswered in this study. Perhaps a way of responding had been learned, or a transfer of a set which temporarily influences the responses made. We might presume that some implicit attitude toward what the subject thought it was that he was to do may have been transferred. However, we can conclude from the results that when homogeneous groups with no previous experience with the Rorschach test are given different types of perceptual training for a short period of time, in a situation separate from the Rorschach itself, significant differences result in their tendency to produce *W* responses and to give organized responses. Thus, the groups present rather different pictures in their Rorschach records. The Gottschaldt group seems to possess a moderate amount of intelligence of a practical sort, while the Mooney group and the control group appear to possess a higher, more abstract type of intelligence. Apparently it was easier to eliminate tendencies to organize and to give *W*'s, than it was to increase this tendency. Perhaps training to increase the *W* and *Z* scores must be accompanied by instructions to see *W*'s, or by an increasing motivation of subjects to give *W*'s by increasing their ego-involvement in the task.

The facility with which the *W* and *Z* scores were altered indicates a need for caution in interpreting these scores in the light of hypotheses advanced for them in Rorschach theory. Perhaps greater structuring of the test situation might lessen the influence of extraneous factors on the protocol. However, Rorschach enthusiasts may defend the test on grounds that these changes in records are superficial and occurred in an unusual setting. They might claim that when a record is interpreted as a whole, by a skilled clinician, an accurate picture of the personality will emerge despite changes in a few scores. Clinical examination of the protocols, as well as the statistics of this study, impress one with the differences between the responses given by the various groups.

If a test is very susceptible to situational influence, then for a clinician to use it and to make accurate predictions for a particular subject will require years of training at skills that are imperfect and never readily communicated to others. Objective personality variables need to be more precisely defined and more accurately measured, with increased effort at reducing the intuitive aspects of the measuring instruments. Only by this refinement will better predictions of behavior be made available.

Summary

The purpose of this study was to investigate some variables which influence the organization of Rorschach responses and response to whole blot areas. Forty-nine volunteer college males were randomly assigned into one of three groups. Two groups received training on perceptual tasks, either the Mooney Closure Figures, or the Gottschaldt figures, before taking the Rorschach. The control group was administered the Shipley–Institute of Living Scale for Measuring Intellectual Impairment, a verbal and nonperceptual test, during its training period. Following the training and a short recess, another experimenter administered the Rorschach test, which was introduced as a second experiment. Since W and Z scores depend on the number of responses given, length of record was controlled by analyzing only the first response made to each card. Four simple analyses of variance were made of W and Z scores. Values of F were statistically significant and tests for the significance of the differences between groups were made.

1. The group trained on the Gottschaldt figures had significantly lower W and Z scores than the other two groups.

2. While the Mooney-trained group's W and Z scores were higher than the control groups, as predicted, these differences were statistically insignificant.

So long as situational variables incidental to the testing situation can be suspected of creating a significant alteration in the W and Z scores, then the variance of these scores that is attributable to pervasive personality dimensions is in serious question. Thus, further caution must be exercised in the interpretation of Rorschach W and Z scores in light of the Rorschach hypotheses given for them. It is suggested that greater structuring of the test situation might decrease the tendency of W and Z scores to reflect transient sets.

Received July 7, 1955.

References

1. Abramson, L. S. The influence of set for area on the Rorschach test results. *J. consult. Psychol.*, 1951, **15**, 337–342.
2. Beck, S. J. *Rorschach's test: I. Basic processes.* (2nd Ed.) New York: Grune & Stratton, 1950.
3. Beck, S. J. *Rorschach's test: II. A variety of personality pictures.* New York: Grune & Stratton, 1945.
4. Bergmann, G. Theoretical psychology. *Annu. Rev. Psychol.*, 1953, **4**, 435–458.
5. Carp, A. L., & Shavzin, A. R. The susceptibility to falsification of the Rorschach psychodiagnostic technique. *J. consult. Psychol.*, 1950, **14**, 230–233.
6. Cronbach, L. J. Statistical methods applied to Rorschach scores: A review. *Psychol. Bull.*, 1949, **46**, 393–429.
7. Eichler, R. M. Experimental stress and alleged Rorschach indices of anxiety. *J. abnorm. soc. Psychol.*, 1951, **46**, 344–355.
8. Eichler, R. M. Some comments on the controlling of differences in responses on the Rorschach test. *Psychol. Bull.*, 1951, **48**, 257–259.
9. Fosberg, I. A. Rorschach reactions under varied instructions. *Rorschach Res. Exch.*, 1938–39, **3**, 12–35.
10. Fosberg, I. A. An experimental study of the reliability of the Rorschach psychodiagnostic technique. *Rorschach Res. Exch.*, 1941, **5**, 72–88.
11. Harrower-Erickson, M. R., & Steiner, M. E. Modification of the Rorschach method for use as a group test. *J. genet. Psychol.*, 1943, **62**, 119–133.
12. Hutt, M. L., Gibby, R., Milton, E. O., & Pottharst, K. The effect of varied experimental

"sets" upon Rorschach test performances. *J. proj. Tech.*, 1950, **14**, 181–187.

13. Keyes, E. J. An experimental investigation of some sources of variance in the whole response to the Rorschach ink blots. *J. clin. Psychol.*, 1954, **10**, 155–160.
14. Kimble, G. A. Social influence on Rorschach records. *J. abnorm. soc. Psychol.*, 1945, **40**, 89–93.
15. Lord, Edith E. Experimentally induced variations in Rorschach performance. *Psychol. Monogr.*, 1950, **64**, No. 10 (Whole No. 316).
16. Mooney, C. M. *Closure test: Instructions for administration and scoring.* Montreal, Canada: Department of Psychology, McGill Univer.
17. Norman, R. D., Liverant, S., & Redlo, Miriam. The influence of a superficial immediately preceding "set" upon responses to the Rorschach. *J. consult. Psychol.*, 1952, **16**, 261–264.
18. Rorschach, H. *Psychodiagnostics.* New York: Grune & Stratton, 1942.
19. Shipley, W. C. *Shipley–Institute of Living Scale for Measuring Intellectual Impairment.* Hartford, Conn.: Institute of Living, 1946.
20. Thurstone, L. L. A factorial study of perception. *Psychometr. Monogr.*, 1944, No. 4.
21. Thurstone, L. L. *Concealed figures: Form A.* Chapel Hill, N. C.: Psychometr. Lab., Univer. of North Carolina, 1951.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Rorschach Summary Scores in Differential Diagnosis[1]

Irwin J. Knopf
State University of Iowa[2]

The discrepancy between reports of clinical experience and research data has been a serious dilemma in Rorschach circles for a number of years. On the one hand, many clinicians have been impressed with the usefulness and the validity of the Rorschach method in a range of applications, while, on the other, research findings have not in the main supported this confidence. One such application has been the considerable use of the Rorschach as an aid in differentiating psychiatric disorders.

Formulated on the assumption that differences in psychiatric conditions would be reflected in Rorschach data, early investigations were primarily concerned with descriptive reports of the typical Rorschach performance of one or more psychiatric populations. As a result, statistical and/or experimental controls were not generally employed, but instead, clinical description was accepted without more rigorous verification. Later, some workers attempted to isolate the qualitative differences purported between patient groups and yet preserve the "holistic" nature of the test findings by deriving patterns or signs which collectively seemed to discriminate among nosological groups. Thus, a variety of signs were reported, for example, signs which were found in brain-damaged individuals, in psychoneurotics, in schizophrenics, and which were useful as indices of adjustment in the evaluation of psychotherapy (17, 18, 14, 8, 22, 15). However, when many of these signs were employed in subsequent investigations, the significant discriminations reported earlier were not corroborated (16, 3, 21, 7, 11). The fact that signs derived in the original investigations provided better discrimination within the initial sample than within subsequent populations is not too surprising. In some of these studies, signs were obtained by selecting for prediction those few aspects of Rorschach performance which showed the highest relationships from among the available predictors which had a low correlation with the criterion. Such a procedure tends to capitalize on chance fluctuations within the sample, and consequently may result in a spuriously high multiple-correlation coefficient. When signs which were derived in this manner are applied to subsequent populations, it can be expected that the coefficient will be of lower magnitude and of less predictive value than that obtained in the parent population.

Other investigators have studied the extent to which Rorschach single summary scores can discriminate among psychiatric groups. Wittenborn and Holzberg (23) studied 39 summary scores with five patient groups and found that one score (*CF*) significantly discriminated the manic from the depressed patients. Friedman (6) evaluated the discriminative effectiveness of Rorschach scores with two groups of normal adults and 30 schizophrenics. He found eight Rorschach variables which significantly differentiated the schizophrenics from both groups of normal subjects. Reiman (19) evaluated 86 scores with replicated samples of ambulatory schizophrenics and neurotics. The results indicate that six scores were significant at the .10 level of confidence or better between the clinical groups for both samples. Kobler and Steil (12) reported no statistical differences of any consequence between the Rorschach scores of the paranoid and depressive subgroups of involutional melancholic patients.

---

[1] A portion of this paper was presented at the American Psychological Association meetings in New York, 1954.

[2] From the Dept. of Psychiatry, College of Medicine, and the Division of Clinical Psychology, Iowa Psychopathic Hospital.

While the findings with respect to single summary scores are predominantly negative, unequivocal evaluation of these data is difficult. Most studies were able to obtain a few scores which discriminated among psychiatric groups. However, in some instances, the positive findings could be expected by chance because of the large number of tests of significance computed. It should also be noted that although a variety of Rorschach scores have been reported as differentiating diagnostic groups, there has not been a great deal of consistency for these scores to appear repeatedly as diagnostic from study to study. In addition, certain methodological limitations such as small samples, incomplete statistical or experimental controls, and vaguely defined diagnostic criteria have complicated the interpretation of the results.

In the light of the ambiguous nature of the research findings, and the extensive application of the instrument, the need for a systematic evaluation of the Rorschach as a psychodiagnostic technique seems indicated. An investigative program of this sort is under way at the Iowa Psychopathic Hospital. The present study represents one major phase of this research program, and it specifically deals with the problem of determining the extent to which Rorschach summary scores can differentiate psychiatric groups.

Procedure

Our subjects were selected on the basis of the following criteria: (*a*) chronological age of 15 years or older; (*b*) unanimous agreement among psychiatrists as to diagnosis both on admission to and discharge from the hospital; (*c*) diagnosis was restricted to psychoneurosis (*Pn*), psychopathy (*Pp*), or schizophrenia (*Sc*); (*d*) diagnosis was independent of the Rorschach data,[3] and (*e*) the number of Rorschach responses (*R*) would not contribute to a significant difference in the mean number or the variance of responses for the three groups.

Initially, over 800 case records of patients who were 15 years or older, and who were given the Rorschach test during the six-year period from 1948 to 1953, were examined in order to check on the agreement and consistency of the psychiatric diagnosis. Each case folder contained an initial diagnostic impression usually made by a psychiatric resident or a staff psychiatrist, an admission and staff diagnosis, and a final discharge staff diagnosis, both of which were made by several residents and one or more staff psychiatrists. In this way 339 Rorschach records were obtained, and the number of responses per record was determined. Statistical tests showed no significant difference between the mean number of responses for the three groups, while the variances between the groups were significantly different. Inspection of the data indicated that two cases in the *Pn* group, each with 153 responses, were contributing greatly to the heterogeneity of variance. Consequently they were withdrawn from the sample and statistical tests of means and variances were recomputed. The groups showed no differences in either the total number of responses or variances, so that the effects of *R* on other Rorschach scores for the three groups were considered approximately equal.

A total of 337 Rorschach protocols meeting all the criteria and obtained from 131 *Pn*'s, 106 *Pp*'s, and 100 *Sc*'s comprised the basic data for this study. In order to assure equal treatment of the data for all subjects, each protocol was rescored according to the scoring system described by Hertz (9).[4] The incidence of clinical types which were included within each of the three major diagnostic groups is presented in Table 1. Additional subject characteristics of the three groups are given in Table 2. From this it will be noted that there was an unequal number of males and females in each group, and that this was most discrepant in the *Pp*'s. The small differences in age between the groups were not statistically reliable, although the differences in education and length of hospitalization

---

[3] It was not possible to obtain complete independence of diagnosis and Rorschach data. However, this criterion was met to the extent that the initial impression and the admission staff diagnosis were made prior to the Rorschach administration.

[4] The author wishes to express his appreciation to Donald Spangler for rescoring each Rorschach protocol.

Table 1

Frequencies of Subclinical Types Within Each Major Diagnostic Group

Psychoneurotics		Psychopaths		Schizophrenics	
Mixed	31	Psychopathic pers.	38	Paranoid	46
Anxiety	30	Path. sexuality	18	Simple	13
Psychasthenia	28	Path. emotionality	14	Hebephrenic	8
Hysteria	22	Psychotic episodes	12	Mixed	7
Neurasthenia	5	Neurotic traits	8	Catatonic	4
Impulse neurosis	5	Inadequate type	7	Acute	3
Hypochondriasis	4	Asocial trends	3	Defect state	1
Psychosomatic	2	Alcoholism	2	Unclassified	18
Reactive depress.	2	Paranoid tendencies	2		
Unclassified	2	Exhibitionism	1		
		Malingering	1		
Total	131		106		100

were significant at the .01 level of confidence. The educational level of the *Sc*'s was slightly higher than that of the *Pn*'s and *Pp*'s, while the length of hospitalization was slightly shorter for the *Pp*'s than for the other two groups. Data were also available with regard to the subject's previous admission and illness. Thirty-seven per cent of the *Pn* group had been ill prior to this admission, whereas 24 per cent of the *Pp*'s and 23 per cent of the *Sc*'s had been ill previously. Generally these figures indicate that approximately 72 per cent of the total sample were first admissions at the time of the Rorschach administration, and that their condition was not, for the most part, considered chronic.

Medians, means, and standard deviations were computed for each clinical group on the following summary scores: *R*, *W*%, *D*%, *Dr*%, *S*, *F*%, *F* + % of 80, *F* + % of 60, *M*, *FM*, *Fm*, *FC*, *CF*, *C*, *Fch*, *chF*, *ch*, *Fch′*, *ch′F*, *ch′*, *Fc*, *cF*, *c*, *A*%, *Ad*, *H*%, *Hd*, Nature, Blood, Sex, Anatomy, Object, Fire, Position, Contamination, *P*, *O*%, and Rejection. Inspection of these figures indicated marked differences in the medians and the means for many of the scores, supporting previous observations that most Rorschach scores are not distributed normally (4, 5). In addition almost half of the scoring categories had medians of zero and means of less than 1.0, which not only suggests that these scores occurred very rarely, but also that they have very limited utility in individual differential diagnosis. With the exception of the *c*, *ch′*, Position, and Contamination scores which occurred very infrequently, chi-square tests of independence were employed to evaluate the significance of differences between the groups for the remaining 34 summary scores. The over-all median derived from the total sample (337) for each score was used as the empirical cutoff point. Yates's correction for continuity was applied to the data wherever

Table 2

Subject Characteristics for the Three Groups ($N = 337$)

	Psychoneurotics (56 M, 75 F)		Psychopaths (79 M, 27 F)		Schizophrenics (57 M, 43 F)	
Sex	Mean	*SD*	Mean	*SD*	Mean	*SD*
Age	27.3	8.6	26.9	9.8	27.7	9.6
Educ.	11.7	2.5	11.1	2.5	12.4	3.2
Length of hosp. (days)	58.1	43.7	40.5	37.5	55.8	45.3

Table 3

Medians, Means, and Sigmas for the Larger Clinical Groups on the Significant Scores and the Probability Values Obtained from Intergroup Comparisons

	Pn (*N*=131)			*Pp* (*N*=106)			*Sc* (*N*=100)			Intergroup chi-square value probabilities*		
Score	Mdn	Mean	*SD*	Mdn	Mean	*SD*	Mdn	Mean	*SD*	*Pn-Pp*	*Pp-Sc*	*Pn-Sc*
Dr%	11.0	14.4	13.8	6.5	10.0	11.4	15.0	15.8	12.8		.01	
A%	50.0	48.0	17.4	51.0	50.2	18.0	44.0	43.4	17.0		.01	
FM	2.0	2.7	2.4	2.0	2.6	2.5	2.0	2.0	2.0			.02
P	5.0	5.1	2.2	5.0	5.2	2.5	4.0	4.1	2.0		.01	.02
Sex	0.0	0.5	1.2	0.0	0.2	1.0	0.0	0.9	2.8	.01	.01	
Anatomy	1.0	2.5	3.3	1.0	1.7	2.4	1.0	2.6	4.3	.01	.01	

* Only probability values at or beyond the .05 levels are reported.

cell frequencies were lower than ten (13). The null hypothesis was retained with those chi-square values which did not meet the minimum requirement of the .05 level of confidence.

Results

Although tests of significance were not computed, the incidence of occurrence of contaminated and position responses will be presented because some Rorschach workers have regarded these responses as diagnostically important in that they are almost always associated with schizophrenia or psychosis (1, 2, 10, 20). Our data, however, indicate that these responses can and do occur in other psychiatric conditions. For example, we found contaminated responses in 9 *Sc*'s, 2 *Pn*'s, and 1 *Pp*, and position responses in 2 *Sc*'s, 5 *Pn*'s, and 1 *Pp*. Moreover, when we consider that only 12 patients out of the total sample of 337 produced contaminated responses and only 8 patients produced position responses, it is apparent that diagnostic classification cannot be effectively made solely on the basis of the presence or absence of these responses.

The over-all chi-square tests applied to each Rorschach score for the three clinical groups resulted in significant values for the following six scores: *Dr%*, *A%*, *FM*, *P*, Sex, and Anatomy. Three additional chi-square values were computed for each of these significant scores to determine more specifically which group or groups the scores discriminated among. Table 3 lists the medians, means, and standard deviations for these scores, as well as the probability level obtained from the separate comparisons of the clinical groups. Most apparent from this table is the failure of any one score to discriminate among all three groups, although three scores were discriminative in two comparisons. It will also be noted that five scores significantly differentiated the *Pp*'s from the *Sc*'s, whereas only two scores differentiated the *Pn*'s from the *Pp*'s, and the *Pn*'s from the *Sc*'s. Inspection of the data and reference to the medians and means listed in Table 3 indicates the direction of significance. There were more *Pp*'s who were lower on *Dr%*, and higher on *FM* than *Sc*'s; more *Pn*'s who were lower on *Dr%*, and higher on *A%* than *Sc*'s; more *Pn*'s who were higher on *FM* than *Sc*'s; more *Pp*'s and *Pn*'s who were higher on *P* than *Sc*'s; and more *Pn*'s and *Sc*'s who were higher on Sex and Anatomy responses than *Pp*'s.

The *F* + % score is often regarded as important in differentiating psychotic from nonpsychotic subjects. However, the over-all median value of 80 per cent which was obtained from the total sample did not discriminate among the clinical groups. Beck (2) has suggested 60 per cent as a diagnostically useful cutoff point, and consequently this score was employed with the present data. The chi-square values indicated significant differences between the *Sc*'s and the *Pn*'s, and the *Pp*'s and *Pn*'s at the .001 and .05 levels of confidence respectively, while no significant difference was obtained between the *Pp*'s and the *Sc*'s. Recognizing that there are differ-

Table 4

Medians, Means, and Sigmas for the Clinical Groups of 50 Cases on the Significant Scores and the Probability Values Obtained from Intergroup Comparisons

	Pn ($N=50$)			*Pp* ($N=50$)			*Sc* ($N=50$)			Intergroup chi-square value probabilities*		
Score	Mdn	Mean	*SD*	Mdn	Mean	*SD*	Mdn	Mean	*SD*	*Pn-Pp*	*Pp-Sc*	*Pn-Sc*
Dr%	8.0	12.7	13.0	7.0	9.9	11.1	14.0	16.5	13.2		.01	.05
M	1.0	1.8	2.2	1.0	1.8	1.5	1.0	1.6	2.8		.01	
Fch'	1.0	1.1	1.5	0.0	0.5	1.3	0.0	0.5	1.0	.01		.05
P	5.0	4.9	2.2	5.0	5.2	2.4	4.0	4.0	2.0		.01	
Sex	0.0	0.6	1.6	0.0	0.1	0.4	0.0	1.0	2.5		.02	
Anatomy	1.0	2.6	3.2	1.0	1.5	1.9	2.0	2.8	2.9	.05	.01	

* Only probability values at or beyond the .05 levels are reported.

ences in the Hertz and Beck scoring systems which include *F* + tables and procedures for computing *F* + %, these findings nevertheless suggest that Beck's empirical cutoff point of 60 per cent may also be discriminative with Hertz's scoring method.

In evaluating the results, it seemed important to consider the extent to which chance factors could account for the significance of differences between the groups for the seven Rorschach scores. Having computed 34 overall chi-square values, we can expect approximately two to be significant merely by chance at the .05 level of confidence. In the light of this possibility, a more rigorous examination of the stability of the present findings seemed essential. Therefore, 150 cases, 50 from each clinical group, were randomly selected from the total parent sample of 337. Statistical treatment and the analysis of the data for this sample was the same as that previously described for the parent population. Chi-square values were obtained separately for the 34 Rorschach scores, and only those scores which were significant on both samples were regarded as stable.

The analysis of the data for the new sample revealed that the *Dr*%, *M*, *Fch'*, *P*, Sex, and Anatomy scores significantly differentiated the groups at the .05 level of confidence or better. The medians, means, and standard deviations for each score together with the probabiilty levels obtained from the separate comparisons of the groups are shown in Table 4. These results indicate the *Pp*'s and the *Sc*'s were significantly differentiated by five scores, whereas only two scores differentiated the *Pn*'s from the *Pp*'s, and the *Pn*'s from the *Sc*'s. In comparing these findings with those obtained from the parent sample, we find, that only the *Dr*%, *P*, Sex, and Anatomy scores significantly discriminated the clinical groups on both samples, and thus met the criterion of stability. While all four stable scores discriminated the *Pp*'s from the *Sc*'s, these scores were less sensitive in differentiating the *Pn*'s from the *Sc*'s, and the *Pn*'s from the *Pp*'s in that only one score for each comparison was found to be significant with both samples (*Dr*% and Sex, respectively).

Summary and Conclusions

To determine the extent to which Rorschach summary scores could discriminate among psychiatric populations, a total of 337 carefully selected Rorschach records obtained from 131 *Pn*'s, 106 *Pp*'s, and 100 *Sc*'s were analyzed. Chi-square tests of independence were computed on 34 Rorschach summary scores for the total sample, and also for a second sample of 150 cases drawn randomly from the parent sample. Only scores which were discriminative on both samples were considered stable. The results showed that:

1. Most Rorschach summary scores are not normally distributed.

2. Almost half of the scoring categories had medians of zero and means of less than 1.0, not only indicating the rareness of these responses but also underscoring the limited utility of these scores for differential diagnosis.

3. Contaminated and position responses can and do occur, albeit infrequently, in all three groups and cannot be regarded as pathognomonic of psychosis or schizophrenia.

4. On an over-all basis, four scores: *Dr%*, *P*, Sex, and Anatomy significantly discriminated among the groups on both samples at or beyond the .05 level of confidence.

5. When specific tests of significance were made, no single summary score significantly differentiated all three clinical groups.

6. For practical purposes, Rorschach summary scores cannot be regarded as effective in differentiating psychiatric groups.

Received July 19, 1955.

References

1. Beck, S. J. *Rorschach's test: I. Basic processes.* (2nd Ed.) New York: Grune & Stratton, 1950.
2. Beck, S. J. *Rorschach's test: II. A variety of personality pictures.* New York: Grune & Stratton, 1947.
3. Cronbach, L. J. Statistical methods applied to Rorschach scores: A review. *Psychol. Bull.*, 1949, **46**, 393–431.
4. Berkowitz, M., & Levine, J. Rorschach scoring categories as diagnostic "signs." *J. consult. Psychol.*, 1953, **17**, 110–112.
5. Fiske, D. W., & Baughman, E. E. Relationships between Rorschach scoring categories and the total number of responses. *J. abnorm. soc. Psychol.*, 1953, **48**, 25–32.
6. Freidman, H. A comparison of a group of hebephrenic and catatonic schizophrenics with 2 groups of normal adults by means of certain variables of the Rorschach test. *J. proj. Tech.*, 1952, **16**, 352–360.
7. Hamlin, R. M., Albee, G. W., & Leland, E. M. Objective Rorschach "signs" for groups of normal, maladjusted, and neuropsychiatric subjects. *J. consult. Psychol.*, 1950, **14**, 276–282.
8. Harrower-Erickson, Molly R. The values and limitations of the so-called "neurotic signs." *Rorschach Res. Exch.*, 1942, **6**, 109–114.
9. Hertz, Marguerite R. Scoring the Rorschach inkblot test. *J. genet. Psychol.*, 1938, **52**, 15–64.
10. Klopfer, B., & Kelley, D. M. *The Rorschach technique.* Yonkers, N. Y.: World Book Co., 1946.
11. Knopf, I. J. The Rorschach test and psychotherapy. *Amer. J. Orthopsychiat.*, 1956, **26**, in press.
12. Kobler, F. J., & Steil, Agnes. The use of the Rorschach in involutional melancholia. *J. consult. Psychol.*, 1953, **17**, 365–370.
13. Lewis, D., & Burke, C. J. The use and misuse of the chi-square test. *Psychol. Bull.*, 1949, **46**, 433–489.
14. Miale, Florence, & Harrower-Erickson, Molly R. Personality structure in psychoneuroses. *Rorschach Res. Exch.*, 1940, **4**, 71–74.
15. Muench, G. A. An evaluation of nondirective psychotherapy by means of the Rorschach and other indices. *Appl. Psychol. Monogr.*, 1947, No. 13.
16. Nadel, A. B. A qualitative analysis of behavior following cerebral lesions. *Arch. Psychol., N. Y.*, 1938, **32**, No. 224.
17. Piotrowski, Z. A. On the Rorschach method and its application in organic disturbances of the central nervous system. *Rorschach Res. Exch.*, 1936–37, **1**, 23–40.
18. Piotrowski, Z. A. The Rorschach ink-blot method in organic disturbances of the central nervous system. *J. nerv. ment. Dis.*, 1937, **86**, 525–537.
19. Reiman, G. W. The effectiveness of Rorschach elements in the discrimination between neurotic and ambulatory schizophrenics. *J. consult. Psychol.*, 1953, **17**, 25–31.
20. Rorschach, H. *Psychodiagnostics.* New York: Grune & Stratton, 1951.
21. Rubin, H., & Lonstein, M. A cross validation of suggested Rorschach patterns associated with schizophrenia. *J. consult. Psychol.*, 1953, **17**, 371–372.
22. Thiesen, J. W. A pattern analysis of structural characteristics of the Rorschach test in schizophrenia. *J. consult. Psychol.*, 1952, **16**, 365–370.
23. Wittenborn, J. R., & Holzberg, J. D. The Rorschach and descriptive diagnosis. *J. consult. Psychol.*, 1951, **15**, 460–463.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Affective Ratio and 8-9-10 Per Cent on the Rorschach Test for Normals and Psychiatric Groups[1]

Earl S. Taulbee
VA Mental Hygiene Clinic, Omaha, and University of Nebraska College of Medicine

Boyd D. Sisson
Veterans Administration Hospital, Omaha, Nebraska

and Charles O. Gaston
University of Houston

Analysis of the Rorschach test by either the Klopfer or Beck scoring systems includes a comparison of the response productivity for those cards which are entirely chromatic (8–10) with the number of responses for the remaining cards. This factor in the Klopfer system, the "8–9–10 per cent" (8–10%), is the percentage contributed by Cards 8, 9, and 10 to the total number of responses. The standard for this factor is based on statistical estimations according to the assumption of *D* responsiveness and the distribution of *D* areas in the ten cards which gives an expectancy of 40 per cent of responses in the last three cards. In Beck's system the comparison is an "affective ratio" (*AR*), obtained as a quotient of responses in Cards 8–10 divided by the response total in Cards 1–7. He reports a mean of 0.60 (*SD* 0.19) for a group of normal adults, and considers 0.40 to 0.80 as the normal range. Beck looks on this factor as "a measure of the readiness to quicken to life's pleasurable experiences" (3), and Klopfer treats it as indicating "a responsiveness to stimuli from without which is even less under the conscious control of the subject than the use of action and color elements" (7). Similar significances are assigned to deviations of this factor—when the ratio is high, the subject is volatile, liable to excitement, and stimulated by environmental impact; when low, he is under-responsive to emotion-toned stimuli and inhibited or lacking in responsiveness to strong environmental impact.

The literature discloses several studies dealing with the validity of color as the determinant of responsiveness in the 8–10% hypothesis. Sappenfield and Buker (9), Dubrovner, Von Lackem, and Jost (5), Perlman (8), Buker and Williams (4), Allen, Stiff, and Rosenzweig (1), using various patient and normal groups, and all employing the technique of achromatic presentations, agreed in finding that the number of responses on 8–9–10 is not a function of color. It was pointed out (8), however, that while the experimental evidence indicates that color is not an active factor in the 8–10%, the possibility remains that there is validity in this factor as a personality variable.

The purposes of the present study are, first, to explore the validity of the 8–10% and the *AR* as a gross personality variable insofar as personality is implied in normal and psychodiagnostic groups, and second, to supply some normative data for the 8–10% hypothesis, and to supplement the norms for the affective ratio.

Study

The subject groups consisted of normal, neurotic, and schizophrenic adults, and normal, behavior problem, and mentally defective children. The neurotics and schizophren-

---

[1] From the Veterans Administration Hospital, Omaha, Nebraska.

Table 1

Subject Groups, Age, Response Total, *AR*, and 8–10%

Group	*N*	Age		*R*		*AR*		8–10%	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Adults									
Normal	181	39.31	10.89	28.36	16.24	.650	.238	38.2	8.1
Neurotic	70	32.96	6.26	22.67	10.71	.594	.292	35.4	10.5
Schizophrenic	119	29.81	6.21	27.24	18.75	.523	.245	32.7	10.5
Children									
Normal	196	12.97	3.14	27.20	12.29	.643	.248	37.8	8.9
Feebleminded	39	12.95	2.09	14.77	3.34	.666	.299	38.4	9.3
Behavior problem	49	12.37	3.02	20.67	10.14	.567	.177	35.4	7.0

ics were VA hospital inpatients diagnosed by the psychiatric staff independent of consideration of the Rorschach test 8–10% or *AR;* the behavior problem group was composed of children referred to a psychological service by schools and courts for behavior disturbances; the mental defectives were institutionalized; the normal subjects were residents of a small midwestern community[2] located in the same geographical locale from which the diagnostic groups came.

The Rorschach tests were individually administered and the protocols analyzed for computation of the 8–10% and affective ratio. No record with fewer than 10 responses was included.

Results

The subject groups, identifying information, and findings for the 8–10% and *AR* are presented in Table 1. All groups, excepting the neurotic and schizophrenic, consisted of both males and females, and the data were first analyzed to determine sex differences within each group. Since statistical analysis revealed no conventionally-accepted significant difference in the 8–10% or *AR*, the data for the males and females were combined.

The differences were treated for significance by means of the *t* test and, for the adult groups, the normals and schizophrenics are differentiated significantly by both the *AR* and the 8–10% ($t = 4.44$ and 4.84, respectively, both $p < .001$), and between the normals and neurotics the difference on the 8–10% achieves a level of conventional significance ($t = 2.00$, $p < .05$). It was found, too, that the age differences are significant at less than .01 between all three groups, and the differences in response total (*R*) are significant between the normals and neurotics and between the neurotics and schizophrenics. Thus, further refinement of the data was attempted by selecting 45 subjects in each of the three classifications for the purpose of matching them on mean age and response total. Results for these matched groups are presented in Table 2.

Table 2

Matched Adult Groups, Age, Response Total, *AR*, and 8–10%

Group ($N = 45$)	Age		*R*		*AR*		8–10%	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Normal	35.29	6.86	24.09	11.10	.649	.212	38.4	7.6
Neurotic	34.53	5.76	24.07	11.62	.604	.320	36.8	11.0
Schizophrenic	34.02	4.95	24.16	13.72	.455	.223	29.9	9.4

---

[2] The records were taken from a larger study done under the supervision of Dr. Marshall R. Jones and supported by the University of Nebraska Research Council.

Table 3

Differences on *AR* and 8–10% for the Matched Groups

Groups	*AR* *t*	*AR* *p*	8–10% *t*	8–10% *p*
Normal vs. Schizophrenic	4.22	<.001	4.66	<.001
Neurotic vs. Schizophrenic	2.53	<.02	3.17	<.01

Table 4

Subjects, Cutoff Scores, Number, and Percentage of Cases for Both the *AR* and 8–10%

	Unmatched groups			Matched groups		
Group	*N*	No.cases *AR*⋜.334 8–10%⋜25.	%	*N*	No. cases *AR*.⋜.334 8–10%⋜25.	%
Normal	181	9	4.97	45	1	2.22
Neurotic	70	10	14.29	45	6	13.33
Schizophrenic	119	33	27.73	45	16	35.55

Analysis of these findings reveals that none of the differences between the normals and neurotics is statistically significant at the acceptable levels of confidence. Significant differentiation can be made, however, between the normals and schizophrenics and between the neurotics and schizophrenics on both the *AR* and 8–10% measures, as listed in Table 3.

For the children groups, inspection of Table 1 reveals that both the *AR* and 8–10% indices significantly differentiate the Behavior Problem group from both the Normals and Mental Defectives, but the differences between the latter two groups are not statistically significant. It is interesting to note the similarity of the mean ratios and their standard deviations for the normal children and normal adults. Further analysis of the data for the children population has not been made.

In the interest of the practical significance of these ratios, a cutoff point was sought which would afford the maximum differentiating effectiveness of the measures. Table 4 presents the number and percentages of cases in each of the groups falling at or below an *AR* of .334 (or the equivalent 8–10%).

All of these differences are significant, and it is felt that these findings support interpretation of the occurrence of a very low *AR* or 8–10% as contributing evidence of pathology of a schizophrenic or neurotic nature. No effective cutoff point could be made at the upper end of the ratio distribution since all three groups showed a similar scatter of a few cases within the upper range. This, it is felt, suggests that a high *AR* or 8–10% is not differentially useful.

Discussion

The results obtained show that the *AR* and 8–10% do have some discriminating value for personality assessment at the level of broad psychodiagnostic classification. Confining attention to the adult populations, it is seen that the normals and schizophrenics differ significantly with respect to mean *AR*, and that the neurotics, more like normals than schizophrenics, are consistent in the direction of deviation from the other two groups. It would probably be agreed that there are personality variables related to normality, neuroticism, and schizophrenia which seem to vary on a continuum of intensity in this order. In terms of personality disturbance, the neurotics as a group, while heterogeneous, are characterized in comparison with normal individuals by the presence of dysfunction, and schizophrenics as even more disordered. Similarly, this view may be applied to the results obtained from the child populations in the sense that a group of behavior problem children differs from their normal peers by disturbances of personality, whereas perhaps the greatest difference between normal and mentally defective children is more exclusively one of intellect than grosser personality functioning, so that these two groups tend to stand together in contrast with behavior problem disorders.

It is felt that the studies which conclude that productivity on Cards 8–10 is not significantly influenced by the presence of color prevent interpretation of these ratios in terms of emotional characteristics of the psychodiagnostic groups or of the traditionally-held relationship of emotional reactivity to color stimuli on the Rorschach test. Findings of this study, however, support Perlman's suggestion that perhaps this factor is related to a personality variable, and supplies the basis for further investigation along such lines. The data reported herein may be of use normatively for both the 8–10% and affective ratio.

Summary

The *AR* and 8–10% ratios were computed from the Rorschach protocols of 654 subjects, representing groups of normal, neurotic, and schizophrenic adults and normal, behavior problem, and mentally defective children. Significant mean differences between some of the groups are obtained which suggest a relationship between the ratios and gross personality differences such as are implied in normal and psychodiagnostic groups. A cut-off point is suggested in the interest of the practical significance of a low ratio in discriminating normal individuals from those with neurotic or schizophrenic pathology. Findings from earlier studies are cited as the basis for caution in interpreting these ratios in terms of emotional reactivity to color stimuli on the Rorschach test.

The present findings are offered as some normative data for the 8–10% and as supplemental norms for the affective ratio.

Received June 27, 1955.

References

1. Allen, R. M., Stiff, Margaret P., & Rosenzweig, M. The role of color in Rorschach's Test: A preliminary survey of neurotic and psychotic groups. *J. clin. Psychol.*, 1953, **11**, 81–83.
2. Beck, S. J. *Rorschach's test*. Vol. II. *A variety of personality pictures*. New York: Grune & Stratton, 1947.
3. Beck, S. J. *Rorschach's test*. Vol. III. *Advances in interpretation*. New York: Grune & Stratton, 1952.
4. Buker, S. L., & Williams, M. Color as a determinant of responsiveness to Rorschach cards in schizophrenia. *J. consult. Psychol.*, 1951, **15**, 196–202.
5. Dubrovner, R. J., Von Lackum, W. J., & Jost, H. A study of the effect of color on productivity and reaction time in the Rorschach test. *J. clin. Psychol.*, 1950, **6**, 331–336.
6. Klopfer, B., Ainsworth, Mary D., Klopfer, W. G., & Holt, R. R. *Developments in the Roschach technique*. Vol. I. *Technique and theory*. Yonkers, N. Y.: World Book Co., 1954.
7. Klopfer, B., & Kelley, D. M. *The Rorschach technique*. Yonkers, N. Y.: World Book Co., 1942.
8. Perlman, Janet A. Color and the validity of the Rorschach 8–9–10 per cent. *J. consult. Psychol.*, 1951, **15**, 122–126.
9. Sappenfield, B. R., & Buker, S. L. Validity of the Rorschach 8–9–10 per cent as an indicator of responsiveness to color. *J. consult. Psychol.*, 1949, **13**, 268–271.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Rorschach Approach Type and the Organization of Cognitive Material[1,2]

E. Robert Sinnett and Ruth Roberts
Student Counseling Bureau, University of Minnesota

The purpose of this study was to test hypotheses about relationships between the way an individual organizes the Rorschach stimuli and the way he thinks in cognitive tasks.

The manner in which a subject organizes his percepts on the Rorschach is commonly interpreted by authorities on the test (e.g., 1, 8, 9) as yielding information about his cognitive processes. A preponderance of whole (*W*) responses indicates that the individual tends to generalize or think in an abstract manner. Emphasis on the usual details (*D*) of the blots signifies that the subject is attentive to the obvious, practical aspects of his environment, and seeing a large proportion of rare details (*Dd*) means that the perceiver may be pedantic and overly concerned with trivial aspects of his environment. Thus, the percentages of *W*, *D*, and *Dd* are used to make inferences about thought processes. These proportions are referred to collectively as the approach type (*Ap*).

Method

Subjects. The subjects were University of Minnesota students who were attending classes in the Educational Skills Clinic at the Student Counseling Bureau. Two samples were drawn: the initial group was composed of 43 students, and the cross-validation group consisted of 69 subjects. Both samples were exhaustive: all students attending classes during the fall quarter and the winter quarter of the 1953–54 academic year participated in the experiment.

Rorschach. A group Rorschach was administered to groups averaging about ten in size. The instructions for the free association were paraphrased from Beck (1) and the inquiry for the location of the responses is described by Harrower and Steiner (5). No inquiry was performed in order to establish the determinants of the responses.

Reading comprehension. After surveying the reading instruments available at the college level, the paragraph reading comprehension section (items 81–100) of the Diagnostic Reading Tests, Survey, Form B, was chosen mainly because the multiple choice answers could be reliably coded in terms of generalizations and details for the purpose of this study.

In addition, the items have alternatives which are "true" in the sense that they are factually presented in the stories, but incorrect in answer to the question. Thus answering a question confronts the subject not only with the problem of selecting an alternative which appears in the paragraph, but he must discriminate the generalization or detail relevant to the question from irrelevant generalizations and details. Unfortunately there were no alternatives which were "true" but irrelevant generalizations and also there were few incorrect generalizations alternatives. The items can be classified into two groups: (*a*) those where the subject must select the relevant generalizations from among alternatives including "true" but irrelevant details, incorrect details, and possibly an incorrect generalization, and (*b*) those where the subject must choose the appropriate detail from

---

[1] This study was supported by a research grant from the Graduate School of the University of Minnesota.

[2] We wish to express our appreciation to Mrs. Dorothy F. Nicholas for suggestions and help in the early phases of this study.

"true" but irrelevant details and incorrect details. Ten items involved selecting the appropriate generalization and seven involved selecting the appropriate detail. Three items were unclassifiable using the criterion of agreement by three of four judges.

Since most of the item alternatives which were generalizations were right answers, a preference for generalizations in answering items would lead to more correct answers to questions directed toward a main idea. Consequently it was predicted that an emphasis on *W* (as indicated by number and percentage of *W*) would be positively related to attaining the correct generalizations on the reading test, but an emphasis on details ($D + Dd$) would be negatively related to attaining correct generalizations on the criterion.

Another index of intellectual organizing activity is Beck's *Z*. This takes into account not only the subject's attempt to organize the stimuli into wholes, but also includes the integration of details commonly perceived as separate. Beck interprets *Z* as ". . . the capacity to grasp relations not perceived by others" (1, Vol. 2, p. 12). Since it has been found that scoring *Z* in unit weights bears such a high relationship to Beck's weighted scores (10), unit scores were used in this study. It was predicted that *Z* would be positively related to attaining correct generalizations and correct details on the reading test.

An overemphasis on *W* in the Rorschach protocol is interpreted as a lack of attention to the practical, obvious aspects of problems. Hence, it was predicted that *W* and *W%* would be inversely related to attaining correct answers to the detail items on the reading test, but an emphasis on details ($D + Dd$) on the Rorschach would be positively associated with attaining correct answers to detail items.[3]

In order to test these hypotheses independent of the effects of reading speed, each of the Rorschach variables used above was related to the percentage of correct generalizations of the generalization items attempted by the subject, and the percentage of correct details of the detail items attempted.[4]

Thus, sixteen tests were made relating the Rorschach variables (*W%*, number of *W* responses, number of $[D + Dd]$ responses, and *Z*) and the reading comprehension variables (number and percentage of correct generalizations, and number and percentage of correct details).

To test the above hypotheses independent of the effects of intelligence, the Vocabulary section included in the reading comprehension test was used for statistical control. Although ACE scores were available on 84 per cent of the sample and it has been more thoroughly studied and used as an intellectual measure, we felt it would have been uneconomical to discard 16 per cent of the sample in order to have complete data for each case. For a sample of 36 cases in the first group the Vocabulary test correlated .59 ($p < .01$) with the ACE. This lends support to the choice of this measure as an index of intelligence.

Story recall. One might argue that the above predictions would not be confirmed because the Rorschach gives a measure of how the individual organizes ambiguous stimulus material, whereas in more structured situations such as the reading comprehension test, the subject may be able to shift his set in accord with the structure provided by the questions. For example, if a subject with a high *W%* were asked a question involving the discrimination of relevant from irrelevant details, he might shift in accord with the set provided by the question. *W%* itself is amenable to shift as a function of the set given by the experimenter (3).

For this reason another criterion task was selected to see if the Rorschach variables would predict performance in a less structured task. For this purpose free recall of stories containing both generalizations and details was chosen. The stories used in this study may be found in Center and Persons (2, pp. 28–30). Each story contained content which could be coded into two generalizations and two details. After reading each story the subjects were told to write a brief summary of the material they had just read.

---

[3] Originally, we had planned to use number and percentage of *D* and *Dd*, but the number of *Dd* responses was too few to warrant a separate analysis ($M = 1.25$). This finding is comparable to that of Harrower-Erickson and Steiner (5) who obtained a mean of 1.3 *Dd* responses.

[4] These analyses were suggested by Dr. Bernard S. Aaronson.

The hypotheses relating Rorschach variables and this criterion are analogous to those relating Rorschach variables and the previous criterion.

Results

Four judges were used to establish agreement in designating the 80 multiple choice alternatives of the 20-item reading comprehension test as details or generalizations. Each alternative was judged to belong to one of two categories: generalization (*G*) or detail (*D′*). The means of the six interjudge agreements were as follows: *G*, 77 per cent, and *D′*, 72 per cent. All interjudge agreements were significant beyond the one per cent level of confidence, using χ^2 for 1 *df*.

Interjudge agreement on whether or not an alternative was true in terms of adhering to the material presented in the reading paragraphs averaged 79 per cent for the six interjudge comparisons, and all interjudge comparisons were significant beyond the .001 level of confidence using χ^2 for 1 *df*.

The alternatives apparently can be reliably coded into correct and incorrect generalizations and details and also can be coded as true or not true.

For the story recall criterion task, the scoring agreement for two judges ranged from 80 per cent to 95 per cent for scoring the recall of the two generalizations and two details of Story 1 (2, p. 28). The agreement was computed separately for each category vs. all other categories and all the relationships were significant beyond the .001 level of confidence. For Story 2 (2, pp. 29–30) the range of interjudge agreement was from 65 per cent to 78 per cent. One of the generalizations could not be coded reliably, and, for the second, agreement was significant at the .05 level of confidence. Coding of both details was significant beyond the .01 level of confidence. Clearly Story 2 was less satisfactory for scoring agreement. One of the generalizations of Story 2 was excluded from the analysis because interjudge agreement was not statistically significant.

For testing the significance of the relationships between the Rorschach and the criteria, the analysis was done using the tables by Mainland and Murray (7) for the median test. These tables were used as an approximation since one-sided tests seemed appropriate. In addition, for all relationships scatter plots were made to locate optimum cutting points for possible cross validation. Where relationships approached significance, Pearson correlation coefficients were computed to permit statistical control of intelligence by means of partial correlation.

Of the sixteen relationships predicted be-

Table 1

Selected Correlations Between Predictors and Reading Comprehension Variables

	Sample 1		Sample 2	
	r	*p*	*r*	*p*
Rorschach				
Reading comprehension*				
Z vs. no. correct generalizations	.46	.01	.14	n.s.
Z vs. % correct generalizations	.28	.05	.20	.05
W vs. % correct generalizations	.29	.05	.11	n.s.
W vs. % correct details	−.30	.05	.12	n.s.
Vocabulary				
Vocab. vs. no. correct generalizations	.62	.01	.49	.01
Vocab. vs. % correct generalizations	.25	.05	.39	.01
Vocab. vs. correct details	.33	.05	.29	.01
Vocab. vs. % correct details	.31	.05	.23	.05
Vocab. vs. *Z*	.33	.05	.00	n.s.

* % refers to percentage of correct responses to items of a given category attempted by the subject.

Table 2

Selected Partial Correlations Between Rorschach Variables and Reading Comprehension Variables with Vocabulary Held Constant

Reading comprehension*	Sample 1		Sample 2	
	Partial r	p	Partial r†	p
Z vs. no. correct generalizations	.35	.01	—	—
Z vs. % correct generalizations	.22	.10	.21	.05
W vs. % correct generalizations	.27	.05	—	—
W vs. % correct details	−.28	.05	—	—

* % refers to percentage of correct responses to items of a given category attempted by the subject.

† The missing entries were not computed because the zero-order r's were not statistically significant (see Table 1).

tween the Rorschach and the reading comprehension measures, four of the zero-order r's were significant beyond the .05 level of confidence in the expected direction (see Table 1). After partialling out the effects of intelligence (see Table 2), three of these relationships were significant beyond the .05 level of confidence and one was significant beyond the .10 level of confidence. On the cross-validation sample only one of these partial correlations attained statistical significance. Beck's Z is significantly related to the percentage of correct generalizations of the generalization items attempted when Vocabulary is held constant. The partial correlations were small (.22, first sample, and .21, second sample) but statistically significant ($p < .10$ and $p < .05$, respectively).

Vocabulary is related in the expected direction to all four of the reading comprehension criteria beyond the .05 level of confidence in both samples.

For testing the hypotheses about the relationships between the Rorschach variables and the story recall, the story criteria were inspected independently of the Rorschach data for suitable cutting points. For both stories, recall of a generalization vs. no recall of a generalization and recall of one or more details vs. no recall of details, appeared to be meaningful cutting points that were near a median split of the population. Tables for 2 × 2 tests (7) were used as an approximation and where a relationship approached significance Fisher's Exact Test was computed, since one-tailed tests were desired.

In all, twenty tests were made (number of W, $W\%$, number of $[D + Dd]$, Z, and Vocabulary vs. the recall of generalizations and the recall of details for both stories) and none was statistically significant. Therefore the story recall task was omitted in the cross-validation study.

In order to determine if subjects' behavior in the criterion tasks was consistent, the interrelations of the reading comprehension measures and the story recall measures were determined. Of ten tests of significance, three were significant beyond the .05 level of confidence. The number of correct generalizations on the reading comprehension test was significantly positively associated with the recall of generalizations on Story 1 ($\chi^2 = 3.05$, 1, df, $p < .05$ in predicted direction), the percentage of correct generalizations in the reading comprehension test and the recall of generalizations on Story 1 were significantly related ($p < .01$), and the recall of details on Story 1 was significantly positively related to the recall of details on Story 2 ($\chi^2 = 3.40$, $p < .05$ in predicted direction). The absence of some of the expected relationships may be due to the rather low interjudge agreement in scoring the generalization on Story 2.

There is some evidence that the criteria are measuring the same thing, but they are not highly related. This finding is congruent with Goldner's (4) data which showed low but positive interrelations among his measures of approach in problem-solving tasks.

Discussion

In contrast with a study by Goldner (4) there was no evidence of a relationship between Rorschach Ap and approach in other tasks. The difference in outcome may be attributable to the kinds of criterion tasks used. Goldner used cognitive tasks such as the Kohs blocks, anagrams, etc., whereas this study employed meaningful verbal material. Perhaps *Ap* is related to behavior in approaching some kinds of problems, but not to the comprehension and recall of meaningful verbal material. Only *Z* was found to show promise for predicting performance in this domain.

Since *Z* gives more weight to the qualitative aspects of performance than the location categories, perhaps future studies should focus more on the qualitative aspects of Rorschach performance. For research of this sort, it would seem advisable to use the individual Rorschach since a more thorough inquiry is feasible than when using a group technique. Another advantage of using the individually administered record is the greater production of unusual details (6) than in the group technique.

Summary

This study was undertaken in order to test hypotheses about the relationship between the organization of the Rorschach stimuli and the organization of cognitive material. A sample of 43 subjects seeking help for reading and study skills was given the Diagnostic Reading Test, Survey Section, Form B, a group Rorschach, and a free recall task. A cross-validation sample of 69 subjects was given the same battery excluding the free recall task. The findings are as follows:

1. There is no support for the hypotheses relating Rorschach approach and thinking in terms of generalizations or details using either a structured or an unstructured cognitive task.

2. There is evidence that *Z*, a measure of organizational activity on the Rorschach, is related to selecting more highly organized responses in a structured cognitive task when intelligence and speed are controlled.

3. When compared with the Rorschach variables used in this study, vocabulary, an intellectual measure, is a better predictor of choice of relevant details and generalizations from irrelevant generalizations and details regardless of whether speed is controlled.

4. Neither the Rorschach variables nor vocabulary predicted the free recall of generalizations or details from story material.

5. There is some evidence of consistency in the part-whole approach among verbal tasks.

Received June 23, 1955.

References

1. Beck, S. J. *Rorschach's test.* 2 vols. New York: Grune & Stratton, 1945.
2. Center, S. S., & Persons, G. L. *Problems in read-and thinking.* New York: Macmillan, 1941.
3. Gibby, R. G. The stability of certain variables under conditions of experimentally induced sets: I. The intellectual variables. *J. proj. Tech.*, 1951, **15**, 3–26.
4. Goldner, R. H. Individual differences in problem solving behavior. Unpublished doctor's dissertation, Univer. of Chicago, 1952.
5. Harrower-Erickson, Molly R., & Steiner, M. E. *Large scale Rorschach techniques.* Springfield, Ill.: Charles C Thomas, 1945.
6. Hertzman, M. A comparison of the individual and group Rorschach tests. *Rorschach Res. Exch.*, 1942, **6**, 83–108.
7. Mainland, D., & Murray, I. M. Tables for use in fourfold contingency tests. *Science*, 1952, **116**, 591–594.
8. Rapaport, D., Gill, M., & Schafer, R. *Diagnostic psychological testing.* 2 vols. Chicago: Year Book Publishers, 1945.
9. Rorschach, H. *Psychodiagnostics.* Berne, Switzerland: Hans Huber, 1949.
10. Wilson, G. P., Jr., & Blake, R. R. A methodological problem in Beck's organizational concept. *J. consult. Psychol.*, 1950, **14**, 20–24.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

The Relation of Rorschach Indices of Extratension and Introversion to a Measure of Responsiveness to the Immediate Environment[1]

Lester Mann [2]
Camden Area Mental Hygiene Clinic

One of the basic assumptions underlying the Rorschach test's indices of extratension and introversion is that the former are indicative of an orientation toward the environment and the latter of one away from it and toward the self. An accumulating literature has come to support this assumption (6, 8, 10, 12, 13, 14).

The present research is an attempt to contribute further to its validation through the test of a logical corollary. Assuming that extratension is characterized by responsiveness to the immediate environment and introversion by a lack of such responsiveness, the investigation sought to demonstrate relationships between Rorschach E-I indices and an operationally defined measure of responsiveness to the immediate environment.

The following Rorschach E-I indices were selected for study: *M, C Total, C Sum, M:C Total, M:C Sum,* $FM + m$, $Fc + c + C'$, $FM + m:Fc + c + C'$. The first five indices are those originally proposed by Rorschach (11); the latter three were later evolved by Klopfer (7).

The criterion measure of responsiveness to the immediate environment was operationally defined in terms of words referring to the immediate environment written in the Binet association task. The Binet task consists of having a subject write a list of words under experimentally controlled conditions and of determining, through inquiry, their temporal referents. It is assumed that words referring to the immediate environment are indicative of responsiveness to it. It is further assumed that the number of such words offers a measure of the degree of responsiveness. Binet (2), employing the original version of the task with his two daughters, found that Marguerite, clearly an extrovert, gave a large number of references to her immediate surroundings; conversely, Armande, an introvert, gave few such references. For a succinct critical appreciation of Binet's work, the reader is referred to Goodenough's discussion (5). A fuller exposition of the writer's adaptation of the Binet technique has been presented elsewhere, together with a rationale for the acceptance of the criterion measure as an operationally defined measure of extroversion-introversion (9).

The general hypothesis of the study was: Rorschach indices of extratension and introversion are related to the criterion measure of responsiveness to the immediate environment.

The following specific hypotheses were tested:

1. Rorschach indices of extratension, i.e., *C Total, C Sum,* $Fc + c + C'$, are positively related to the criterion measure.

2. Rorschach indices of introversion, i.e.,

---

[1] This article is based on a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, University of North Carolina, 1953. The major part of the investigation was carried out while the writer was a trainee in the VA Clinical Psychology Training Program. The guidance of Drs. Harry W. Crane, Joseph G. Dawson, and Dorothy Terry, members of the writer's thesis committee, and Dr. B. J. Winer, who acted as statistical advisor, is gratefully acknowledged.

[2] Formerly at the Mental Hygiene Clinic of Raleigh and Wake County.

M, $FM + m$, are negatively related to the criterion measure.

3. Rorschach ratio indices of extratension-introversion, i.e., $M{:}C$ *Total*, $M{:}C$ *Sum*, $FM + m{:}Fc + c + C'$, are negatively related to the criterion measure.

Procedure

The experiment was conducted in a small room with a single window through which adjacent buildings and trees could be seen. Noises could be heard from outside the room, both within and without the building. The room was equipped with a table, three chairs, a bookcase holding several objects, and a magazine stand with a few magazines scattered upon it. These articles were its original equipment and, to avoid any possible impression of contrivance, were not altered in any way.

Fifty undergraduate students served as subjects in the experiment which was conducted individually. Two administrations of the Binet association task opened the experimental session. The Rorschach test was then given. A final administration of the Binet task followed. Standard sets of Rorschach inkblots were employed for the test. Sheets of paper, numbered vertically from one through twenty-five, and pen were provided for the Binet task.

The administrative procedure for the Binet task was as follows: The subject was first instructed, "Write twenty-five words, any twenty-five words just as they come to mind. Write no sentences." The task completed, the subject was questioned as to his choice of words. "What did you have in mind when you wrote that word? What made you write it?" If a question remained as to a word's exact temporal referent, the examiner asked whether the subject had "anything particular in mind" when he wrote it.

The Rorschach test was administered in general accordance with the procedure outlined by Klopfer (7). A maximum of ten responses was allowed for each inkblot to insure completion of the test within a reasonable time period.

Beck's system (1) was used to score the determinants in the E-I indices formulated by Rorschach, Klopfer's (7) to score the determinants in the indices he proposed. Certain responses are scored M under the former system but FM or m under the latter. In such cases both M and $FM + m$ indices were credited.

When a percept involved multiple determinants, Beck's rule (1) was followed and multiple credit assigned, i.e., all determinants were equally scored. The one exception taken to the rule was in the circumstance of determinants composing the same index, e.g., only one score would be credited to the index $FM + m$ on the basis of a percept involving both animal and inanimate movements. It was felt that multiple scoring in such cases would overcredit the index.

A total of seventy-five words were obtained through the three administrations of the Binet association task. These were analyzed to determine the number that referred to the immediate environment. The latter constituted the criterion measure.

The sole basis for inclusion of a word within the criterion measure was that of reference to the immediate environment; to any aspect of it and in any manner whatsoever. The immediate environment was defined as extending the entire experimental period, i.e., from the moment the subject entered the experimental room to that of his completing the assignment. Examples of references to it are words referring to the room's furnishings, to sounds coming from without the building, to the subject's physical states at the time of the experiment.

Analysis of the Data

Chi square and tetrachoric r were the statistics employed in the investigation. These counting procedures have been recommended for use with Rorschach data since they are free of assumptions the latter cannot meet, i.e., normality of sample distributions and equality of test units (4).

The χ^2's were computed with corrections for continuity where necessary. One-tailed tests of significance were employed since the directions of relationship were predicted in all instances. The .05 point was considered crucial for rejection of the null hypothesis.

The Thurstone computing diagrams (3) were employed to estimate the r_t's. The z transformation was used to average the latter and to test the significance of differences between them.

The statistical treatment required the rank-order distribution and dichotomization of the experimental variables. The latter was carried out to achieve, as closely as possible, an equal division of the distributions into high and low score groups. If several ties appeared at a distribution median, they were assigned to either group as best approximated equality of division. If but two ties appeared at this point they were randomly assigned, one to each group.

In considering the *total group* of subjects certain of the Rorschach E-I indices were found significantly related to R, i.e., M, C *Total*, $Fm + m$ and $Fc + c + C'$ are related to R beyond the .05 point for the total group. A procedure recommended by Cronbach (4) was employed to control this factor. These indices were studied, as regards their relationship to the criterion measure, within two subgroups formed, on the basis of R, from the total group; a *high* R group, consisting of twenty-four subjects, with R ranging from forty-eight through one hundred, and a *low* R group, comprising twenty-six subjects, with R ranging from twenty-three through forty-six. Their relationships to R are not significant within these subgroups.

In order to obtain over-all estimates of the relationships of M, C *Total*, $FM + m$, and $Fc + c + C'$ to the criterion measure for the total group of subjects, the results obtained from the subgroups were combined, i.e., the χ^2's added, with their significance being determined for two degrees of freedom, and the r_t's averaged through the z transformation. Those E-I indices not significantly related to R in the total group, i.e., C *Sum*, M:C *Total*, M:C *Sum* and $FM + m$:$Fc + c + C'$, were studied directly as regards their relationship to the criterion measure, i.e., in the total

Table 1

Summary of Rorschach and Criterion Measure Data

Variable	Mean	Range	*SD*	Division†	Relation to R	
					χ^2	r_t
		Total group				
R	54.16	23–100	26.63			
M	7.88	0–32	5.98	24–26	2.93*	+.42
C *Total*	7.88	1–18	4.86	24–26	4.16*	+.54
C *Sum*	6.53	.5–17.00	4.31	25–25	1.28	+.31
M:C *Total*	1.76	0–17.00	1.34	25–25	.53	+.19
M:C *Sum*	2.44	0–34.00	4.94	25–25	.04	+.07
$FM+m$	11.7	2–37	6.62	25–25	5.02*	+.54
$Fc+c+C'$	8.04	1–39	6.18	26–24	8.03*	+.73
$FM+m$:$Fc+c+C'$	2.07	.25–8.00	1.59	25–25	.04	+.07
Criterion	22.63	0–66	15.71			
		High R group				
R	77.75	48–100	18.65			
M	10.42	0–32	7.38	12–12	.17	+.25
C *Total*	9.79	1–18	5.20	12–12	—‡	−.25
$FM+m$	7.14	5–37	7.03	11–13	.00	+.02
$Fc+c+C'$	5.38	2–39	5.84	12–12	.00	.00
Criterion	25.25	0–66	16.58			
		Low R group				
R	32.38	23–46	7.42			
M	5.54	1–12	2.65	13–13	.00	+.13
C *Total*	6.12	1–14	3.74	13–13	2.46	+.60
$FM+m$	9.71	2–20	4.49	14–12	.08	+.14
$Fc+c+C'$	3.42	1–13	5.12	14–12	1.07	+.36
Criterion	20.19	2–56	14.45			

† Division of Rorschach E-I variable distributions to determine relationship to R and criterion measure; numbers listed are subject totals in upper and lower halves, respectively.

‡ χ^2 of .17 was in direction opposite to that predicted.

* Significant at or below the .05 point.

Table 2

The Relation of Rorschach E-I Indices to the Criterion Measure

Rorschach E-I Index	Total group			High *R* group			Low *R* group			Combined results of subgroups		
	χ^2	r_t	%† agt.	χ^2	r_t	%† agt.	χ^2	r_t	%† agt.	χ^2	r_t	%† agt.
M				4.17*	−.71	75	.15	−.12	54	6.15*	−.45	64
C Total				1.50	+.48	67	5.54**	+.77	77	10.61**	+.65	72
C Sum	5.12*	+.54	68									
M:*C Total*	5.12*	−.54	68									
M:*C Sum*	2.88*	−.43	64									
$FM+m$				.31	−.24	58	—‡	+.13	46	.45	+.05	52
$Fc+c+C'$				4.17*	+.71	75	—‡	−.13	46	5.67*	+.34	60
$FM+m:Fc+c+C'$	1.28	−.31	60									

Note.—χ^2's for total, high *R*, and low *R* groups are based on one degree of freedom. χ^2's for the combined results of the subgroups are based on two degrees of freedom.

* Significant at the .05 point.

** Significant at the .01 point.

† % of subjects in group whose results are in agreement with hypothesized relationship of Rorschach E-I indices to the criterion measure.

‡ The χ^2's computed for the relationships of $FM+m$ and $Fc+c+C'$ to the criterion measure were in a direction opposite to that predicted. For both indices χ^2 was .08.

group per se, without controls for *R* being instigated.

Table 1 presents a summarization of the data obtained in the total, low *R*, and high *R* groups. The means, ranges, and *SD*'s of the various variable distributions are presented therein, the manner of their division for tests of significance and estimation of correlation, the relationship of Rorschach E-I variables to *R*.

Results and Discussion

The results of the investigation proper are presented in Table 2.

Of the E-I indices studied directly in the total group of subjects, *C Sum*, *M*:*C Total*, and *M*:*C Sum* are related to the criterion measure in statistically significant fashion. The null hypothesis is sustained in the case of $FM + m:Fc + c + C'$.

None of the E-I indices studied in the subgroups are significantly related to the criterion measure in both high *R* and low *R* groups. *M* and $Fc + c + C'$ demonstrate significant relationships in the former but not in the latter, *C Total* in the latter but not in the former, $FM + m$ in neither. However the treatment of these indices in the subgroups was specifically to control for *R*. The combined results, from the subgroups, offering over-all estimates of the relationships to the criterion measure for the total group of subjects, are regarded as the crucial ones for the experiment. They permit rejection of the null hypothesis for *M*, *C Total*, and $Fc + c + C'$ while sustaining it in the case of $FM + m$.

None of the statistically significant results was significantly different from the others.

Summary

A study was carried out with fifty subjects to test the assumption that Rorschach indices of extratension and introversion are indicative of opposed orientations towards the external environment. Assuming that extratension is characterized by responsiveness to the immediate environment and introversion by a lack of such responsiveness, an attempt was made to demonstrate relationships between Rorschach E-I indices and an operationally defined measure of responsiveness to the immediate environment.

The Binet association task furnished the latter measure. The subjects were required to write, on three different occasions, the first twenty-five words that came to mind. The number of these words that referred to the immediate environment formed the criterion measure.

The following Rorschach E-I indices were

found significantly related to the criterion measure at and below the .05 point of significance: M, C *Total*, C *Sum*, $M{:}C$ *Total*, $M{:}C$ *Sum*, $Fc + c + C'$. The null hypothesis was sustained for the relationships of $FM + m$ and $FM + m{:}Fc + c + C'$. None of the statistically significant relationships was significantly different from the others.

Received June 30, 1955.

References

1. Beck, S. J. *Rorschach's test. Vol. 1. Basic processes.* New York: Grune & Stratton, 1944.
2. Binet, A. *L'Etude expérimental de l'intelligence.* Paris: Alfred Costes, 1903.
3. Chesire, L., Saffir, M., & Thurstone, L. L. *Computing diagrams for the tetrachoric correlation coefficient.* Chicago: Univer. Chicago Bookstore, 1933.
4. Cronbach, L. J. Statistical methods applied to Rorschach scores. A review. *Psychol. Bull.*, 1949, **46**, 393–431.
5. Goodenough, Florence L. The appraisal of child personality. *Psychol. Rev.*, 1949, **56**, 123–131.
6. Kimble, G. A. Social influences on Rorschach records. *J. abnorm. soc. Psychol.*, 1945, **40**, 89–93.
7. Klopfer, B., & Kelley, D. M. *The Rorschach technique.* Yonkers, N. Y.: World Book Co., 1942.
8. Lane, Barbara M. A validation test of the Rorschach movement interpretations. *Amer. J. Orthopsychiat.*, 1948, **18**, 292–296.
9. Mann, L., & Branca, A. An operationally defined measure of extroversion-introversion. *J. child Psychiat.*, in press.
10. Meltzoff, J., Singer, J. L., & Korchin, S. J. Motor inhibition and Rorschach movement responses: A test of the sensory-tonic theory. *J. Pers.*, 1953, **21**, 400–410.
11. Rorschach, H. *Psychodiagnostics.* Berne, Switzerland: Hans Huber, 1949.
12. Singer, J. L., Meltzoff, J., & Goldman, G. D. Rorschach movement responses following motor inhibition and hyperactivity. *J. consult. Psychol.*, 1952, **16**, 359–364.
13. Singer, J. L., & Spohn, H. E. Some behavioral correlates of Rorschach's experience-type. *J. consult. Psychol.*, 1954, **18**, 1–9.
14. Werner, H. Perceptual behavior of brain-injured, mentally-deficient children. *Genet. Psychol. Monogr.*, 1945, **31**, 51–110.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Cognitive Inhibition and Rorschach Human Movement Responses[1]

Murray Levine[2]
Veterans Administration Regional Office, Philadelphia, Pa.

and Julian Meltzoff
Veterans Administration Regional Office, Brooklyn, N. Y.

Rorschach (5) implied a triadic relationship among motor behavior, motion perception, and cognitive processes. Although he did not offer a theoretical rationale for his empirical observations, theorists of rather different orientation have postulated similar relationships. Werner and Wapner's (11) sensory-tonic theory of perception has provided a theoretical basis for the vicarious relationship between two members of this triad—motor activity and motion perception. Studying sensory-tonic theory within the framework of the Rorschach situation, previous experiments (2, 7) have revealed the inhibition of a motor act to be followed by an increased production of human movement (M) responses. Further, the hypothesis that individuals who are motorically hyperactive tend to produce fewer movement responses than those who are less active has been supported experimentally (6, 8, 10).

A theoretical link also exists between motor activity and cognition. In psychoanalytic theory, the gradual shift from the primary to the secondary process involves the delay and control of impulse expression, and the substitution of symbolization, fantasy, planning, and thinking as "experimental actions" in place of immediate response on a behavioral level. Rapaport (4) has pointed to the inhibition of activity designed to produce gratification as the primary condition underlying the development of early fantasy and thus thinking.

In a previous paper (1) dealing with cognitive behavior at the level of the formation and inhibition of associations, we have shown that the ability to inhibit an habituated activity at the motor level (writing a phrase as slowly as possible) is directly related to the ability to inhibit learned associations. Those *S*s who were able to slow their writing down for the longer period of time were able to inhibit learned associations and substitute new ones for them more quickly than were poor motor inhibitors. Other studies (6, 9) have shown relationships between the perception of M on the Rorschach, motor inhibition, measures of fantasy on TAT stories, and scores on tasks designed to measure general planfulness. Since motor and cognitive inhibition are interrelated (1), further support of the triadic relationship among motion perception, motor inhibition, and cognitive processes would be obtained were it shown that M is also a measure of the ability to inhibit associations.

The Experiment

Hypothesis. A direct relationship between the production of Rorschach human movement responses and the ability to inhibit associations is hypothesized. It is proposed that *S*s who are sensitive to kinesthetic factors on the Rorschach will be better able to inhibit associations than those who are less perceptive of human movement in the inkblots.

---

[1] From the Veterans Administration Regional Office, Philadelphia, Pa., and Brooklyn, N. Y.

[2] The assistance of Drs. William J. Cohen and Robert Fortier, of the Philadelphia Regional Office, in gathering data is gratefully acknowledged.

clinical testing. From this vantage point, with access to case history material, our observations suggest that the techniques used in this and the previous study do actually reflect the degree to which strong impulses or affects are characteristically controlled in the process of avoiding conflict with society. These impressions, of course, require experimental verification.

Summary and Conclusions

This experiment was designed to test the hypothesis of a direct relationship between the production of human movement (*M*) responses on the Rorschach test and the ability to inhibit associations. Ninety-three *S*s were administered the Rorschach and a task designed to measure the efficiency of cognitive inhibition. The *S*s were divided at the median into High and Low *M* and *M*% groups and these groups were then compared for cognitive inhibition time (CIT). Analysis of the results warrants the conclusion that individuals who are more responsive to kinesthetic stimuli in the Rorschach are better able to inhibit associations than are those who are not productive of *M* responses.

When these findings are considered in conjunction with those of previous related studies, they lend further support to the triadic hypothesis interrelating motor behavior, motion perception, and cognitive processes.

Received August 12, 1955.

References

1. Meltzoff, J., & Levine, M. The relationship between motor and cognitive inhibition. *J. consult. Psychol.*, 1954, **18**, 355–358.
2. Meltzoff, J., Singer, J. L., & Korchin, S. J. Motor inhibition and Rorschach movement responses. *J. Pers.*, 1953, **21**, 400–410.
3. Rapaport, D. *Diagnostic psychological testing.* Vol. 2. Chicago: Year Book Publishers, 1946.
4. Rapaport, D. *Organization and pathology of thought*. New York: Columbia Univer. Press, 1951.
5. Rorschach, H. *Psychodiagnostics*. New York: Grune & Stratton, 1942.
6. Singer, J. L., & Herman, J. Motor and fantasy correlates of Rorschach human movement responses. *J. consult. Psychol.*, 1954, **18**, 325–331.
7. Singer, J. L., Meltzoff, J., & Goldman, G. D. Rorschach movement responses following motor inhibition and hyperactivity. *J. consult. Psychol.*, 1952, **16**, 359–564.
8. Singer, J. L., & Spohn, H. E. Some behavioral correlates of Rorschach's experience type. *J. consult. Psychol.*, 1954, **18**, 1–9.
9. Singer, J. L., & Wilensky, H. Rorschach's movement response as a measure of delaying capacity, fantasy and planfulness. Paper read at East. Psychol. Ass., Philadelphia, April, 1955.
10. Werner, H. Perceptual behavior of brain-injured, mentally deficient children. *Genet. Psychol. Monogr.*, 1954, **31**, 51–110.
11. Werner, H., & Wapner, S. Sensory-tonic field theory of perception. *J. Pers.*, 1949, **18**, 88–107.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Actuarial Prediction

Robert D. Wirt
University of Minnesota

Meehl (2) has pointed out the generally favorable position which statistical prediction appears to have relative to the clinical prediction of behavior. In his search of the literature for pertinent data which could be used to compare the relative power of clinical vs. statistical prediction, he was able to find only a few appropriate studies. The present investigation represents such a comparison.

Data from an earlier study by the author were used (3). In that study it was found that Barron's Ego-Strength scale (1) successfully separated hospitalized psychiatric patients who had shown great improvement in psychotherapy from those who did not improve, as rated by the patients' psychiatrists.

One test of clinical vs. statistical prediction amounts to the difference between running psychometric data through a machine and accepting an actuarial cutting line from which selection for treatment will be made as against running the same test results through the intuitive psyches of one or more clinicians who then make similar selections. In the present study the 19 patients who were rated as the most greatly improved as a result of a short period of psychotherapy from a total N of 535 consecutive admissions to a neuropsychiatric hospital were matched for age, days of hospitalization, and diagnosis with 19 patients who were rated as unimproved by their doctors. The composition of the sample is shown in Table 1. Most of the traditional diagnostic categories were represented, with most cases being classified among the neuroses. As the data in Table 1 suggest, there is a positive relationship between IQ and *Es* score. This is consistent with Barron's finding. The MMPI profiles (scored for the usual four validity and ten clinical scales) for these 38 patients were given to eight expert judges,[1] all experienced in Multiphasic interpretation. They were asked to sort the profiles into two equal groups, one category to include those patients for whom the judge would predict improvement with psychotherapy and the other category for those the judge predicted would not improve in psychotherapy. The results of the judgments are shown in Table 2,

Table 1

Composition of the Sample

Statistic	Greatly improved	Unimproved
Mean age in years	28.5	32.5
Mean days of hospitalization	38.5	39
Mean IQ	124	105
Median *Es* score	50.5	37.5
N	19	19

Table 2

Clinical Sorts of the MMPI Profiles

Judge	Number of correct judgments ($N=38$)	χ^2
A	16	.95
B	22	.95
C	20	.10
D	22	.95
E	26	4.90*
F	20	.10
G	20	.10
H	18	.10

* $p = .04$.

---

[1] The author wishes to express his appreciation to Drs. Harold Gilberstadt, Starke R. Hathaway, Richard C. Kogl, Paul E. Meehl, John J. Regan, Albert Rosen, William Schofield, and Mr. Arthur J. Gallese, who were the judges in this investigation.

where it will be seen that only one judge exceeds chance at an acceptable level of statistical confidence.

The *Es* scale values for these groups misclassify four individuals in each direction, if the median of the total sample is used as the cutting line. Thus the number of correct judgments for the mechanical sorting would be 30, which gives a chi-square value of 12.73, the probability of which is less than .001. In this situation the use of mechanical sorting clearly beats clinical prediction.

Discussion

Service psychologists frequently are asked to make recommendations for treatment. It is a matter of very great personal and economic importance to them and their patients that their predictions be accurate. If results such as those described here are indicative, it might be by far in the best interest of patients to devote our time and skill to the development of refined actuarial and mechanical procedures. However, it should be remembered that what is here called "clinical" is not by any means what most psychologists would mean by that word. Meehl implies that such examples as the one given here can be used to compare the clinical with the statistical approach. As a matter of fact, in the practical situation most psychologists make their recommendations for treatment upon psychometric data, as used in this investigation, together with social history, medical data, interview material, and observation of the patient. An appropriate test of the difference in power between the two approaches would be to have clinical predictions made from total psychological examinations; these to be compared with actuarially derived measures such as the *Es* scale.

A further word should be said about the base-rate problem in prediction. In the present investigation the situation presented to the judges was a gross distortion of the base rates of recovery following psychotherapy in the hospital population used. Of 225 patients treated by psychotherapy alone, only 19 cases were rated greatly improved; but the judges were given the problem of selecting these 19 from an *N* of only 38 cases. The real task of the clinician, or of the machine, in a practical situation would be to identify the good bets for psychotherapy out of the total population. The *Es* scale in the study cited could be stiff competition for a clinical approach.

Received July 26, 1955.

References

1. Barron, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, **17**, 327–333.
2. Meehl, P. E. *Clinical vs. statistical prediction.* Minneapolis: Univer. of Minnesota Press, 1954.
3. Wirt, R. D. Further validation of the ego-strength scale. *J. consult. Psychol.*, 1955, **19**, 444.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Some Determinants of Successful and Unsuccessful Adaptation to Hospital Treatment of Tuberculosis[1]

Louis J. Moran, George W. Fairweather, and Robert B. Morton
Veterans Administration Hospital, Houston, Texas

One large hospital treated and discharged over one thousand veteran and military tuberculous patients in 1952. Both veteran and military patients were treated under the same hospital conditions and by the same medical staff. Of the military patients, 98 per cent were treated successfully (disease rendered inactive). Of the veteran patients, only 33 per cent were treated successfully. This marked difference between the two groups in success of treatment for tuberculosis may be accounted for almost entirely by the voluntary discharge of 47.8 per cent of the veteran group, against the advice of their physicians (6).

This datum typifies what has become the major problem in the successful treatment and control of tuberculosis, that of keeping the patient in the hospital until his treatment is completed. Every year, from one-third to one-half of the patients in tuberculosis hospitals leave the hospital against medical advice (6, 8).

In recent years, a growing number of experimental studies of the problem of "irregular discharge" (leaving the hospital against medical advice) have appeared. Tollen (8) has provided an excellent review of the studies reported prior to 1948. Representative studies since this time have used various measures such as the MMPI (1, 4, 5), tests of intelligence (7), and demographic information (2, 5) in an effort to differentiate the potential "irregular discharge" patient from other hospitalized tuberculous patients.

The present investigation proposes to examine the relationship of the patients' (*a*) family of orientation, (*b*) prehospital life adjustment, (*c*) attitudes towards situations in the hospital, (*d*) behaviors in the hospital, and (*e*) current outside-the-hospital situations, to his success or lack of success in remaining in the hospital until his treatment is completed. For purposes of analysis, it is hypothesized that a high level of adjustment in each of these areas is associated with successful completion of hospital treatment. A formal statement of the hypotheses to be tested is provided in the section on results.

Method

The initial data were obtained from 140 male veterans who were currently receiving treatment for tuberculosis in a large Veterans Administration GM&S hospital. The sample comprised twenty patients from each of seven wards which represented different stages of treatment. Thus, the sample included all individuals receiving treatment for tuberculosis at the installation, with the exception of four who would not cooperate and 40 who were too ill to be seen. Nine months after the initial interview the hospital records were examined to ascertain which patients had left and the nature of their discharge.

The criteria for irregular discharge were: (*a*) the patient must have left voluntarily (*b*) he must have left against the advice of his physician. The criterion for regular discharge was simply that the patient had successfully completed treatment. In addition to these criteria unique to each group, individuals with a history of mixed discharges were eliminated from the sample. Thus, six patients were excluded who had received a discharge in the past that was the opposite of their current discharge. Of the 140 subjects interviewed, 39 met the criteria for regular discharge and 19 met the criteria for irregular discharge within the nine-month period.[2]

---

[1] The authors should like to acknowledge the generous support and guidance of Drs. Daniel E. Jenkins, Hollis G. Boren, and Irving Chofnas, tuberculosis specialists, in the conduct of this study.

Information concerning the patient's past life adjustment and current outside-the-hospital situations was obtained through structured interviews with the patient. The data concerning the patient's adjustment to the hospital encompassed his attitudes toward the institution and his ward behavior. The former was obtained through interview, the latter from behavioral ratings by nurses and aides. Owing to the nature of their illness, each patient was seen for not more than thirty minutes at any one time. The number of such interviews varied from four to seven, depending upon the patient. All responses were recorded on data sheets while the interview was in progress. In the account to follow, a brief description of the content and the method of scoring each measure is provided.[3]

Measurement of Prehospital Life Adjustment

A life adjustment scale was devised, comprising 132 items of information which covered nine areas of living.[4] Here the cardinal principle was to cover comprehensively areas requiring adaptation in the patient's life. The items dealt primarily with factual information rather than with attitudes or judgments. Thus, a patient was not asked about his attitudes toward job, peers, parents, etc., but rather such questions as the number of jobs held, number of close friends, amount of time each parent was home, etc.

Briefly, the areas covered were as follows. *A*. The family background scale, composed of the following five subscales: (*a*) parents' economic status, (*b*) parents' education, (*c*) parents' occupation, (*d*) completeness of family life, i.e., items concerning siblings, parents' marital status, etc., (*e*) security aspects of early life, e.g., who raised the patient, parents' time at home, etc. *B*. The educational adjustment scale consisted of items encompassing grade school and high school truancy, adjustment to peers and teachers, etc. *C*. The social relationship scale covered childhood, adolescent, and adult antisocial behavior, relationships with peers, organizational membership, etc. *D*. The health scale was concerned with childhood, adolescent, and adult illness, amount of medical attention, etc. *E*. The drinking scale was concerned with the amount and type of alcoholic indulgence in adolescence and adulthood. *F*. The religion scale covered childhood, adolescent, and adult church attendance. *G*. The army adjustment scale dealt with relationships with peers and officers, gain in rank, etc. *H*. The occupation adjustment scale was concerned with job level, number of jobs held in past 5 years, reasons for leaving a job, etc. *I*. The marital adjustment scale included marital status, number of marriages, number of children, disciplining of children, etc.

Scoring. After all items of the life adjustment scale had been selected, each possible response to an item was ranked on an a priori basis from least to most adaptive. The score for each possible response to an item was its rank position, with the least adaptive response receiving a score of zero. Thus, the response to an item with two response possibilities received scores of zero and one. An item with three response possibilities received scores of zero, one, two, and so on through the items.

To prevent items with the greatest number of response possibilities from unduly influencing the total scale score, responses to each item were then rescored on a scale from zero to nine. Thus, an item with two responses divided the zero to nine scale into three equal units and the two alternatives received scores of 3 and 6. A three-response item divided the zero to nine scale into four equal units and the three alternatives received scores of 2.3, 4.5, and 6.7, respectively. All items on the scales were so rescored. Any scale score was the sum of the items comprising that scale.

---

[2] A review of the records of this hospital from 1949 to the present indicated that over 75 per cent of all irregular discharges have taken place within nine months after admission. Since the yearly irregular discharge rate at this hospital had been about 50 per cent we expected at the nine month period to have in the irregular discharge sample over 50 of the 140 subjects examined. The occurrence of only 19 irregular discharges in this sample is accounted for in part by a 15 per cent reduction in irregular discharge rate during this period. However, the further reduction of expected cases, not accounted for by the lowered irregular discharge rate, throws light on a hitherto little appreciated aspect of the irregular discharge problem, the implications of which are developed in the Discussion section.

[3] The complete interview schedules and examination forms, together with scoring manuals, have been deposited with the American Documentation Institute. Order Document No. 4741 from the ADI Auxiliary Publication Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $3.75 for photoprints or $2.00 for 35 mm. microfilm. Make checks payable to Chief, Photoduplication Service, Library of Congress.

[4] In the construction of this scale we borrowed ideas freely from many published scales. Our greatest single debt is to Leslie Phillip's Social Attainment Scale (3).

Measurement of Current Hospital Adjustment

Attitude scale. An attitude scale was developed in an attempt to provide a quantitative measure of each patient's attitudinal adjustment to the hospital. The attitude scale comprised 45 items that were selected as constituting important adjustmental problems for tuberculous patients. The problem situations were established by means of extensive unstructured recorded interviews with ten randomly selected tuberculous patients. A compilation of these patient statements constituted the initial set of items. These items were organized, on an a priori basis, into the following areas: (*a*) attitudes about ward regulations, (*b*) attitudes about personnel, (*c*) attitudes about other patients, (*d*) attitudes about hospital facilities.

The 45 items were couched in terms of statements, e.g., "Bed rest is the most important part of treatment." During the interview the statements were read to the patient. The patient was instructed to respond in one of the following four ways: strongly agree, moderately agree, strongly disagree, moderately disagree. The patient's responses were recorded on a data sheet during the interview.

Attitude scale scoring. First, the experimenters made a judgment as to whether agreement or disagreement with a particular item could be considered least or most adaptive to prolonged hospitalization. The response of the patient was then scored as 0, 1, 2, or 3, depending on the intensity of the attitude and its adaptive directionality. Thus, an item considered adaptive in the direction of agreement received the following scores: strongly disagree = 0, moderately disagree = 1, moderately agree = 2, strongly agree = 3. Each item received scores of 0, 1, 2, 3, which were then transformed into the nine-point scale values of 1.8, 3.6, 5.4, and 7.2, respectively. The score for any area was the sum of the items comprising that area.

Behavior rating scale. The guiding motive in the construction of the behavior rating scale was to obtain a comprehensive description of ward behavior. The primary criterion for selection of a specific item was that it described an activity, a plainly observable action, on the part of the patient. In the collection of such items an explicit effort was made to suspend temporarily judgments of the relevancy of the item to prediction of irregular discharge.

Initially, about 80 such items were derived from the knowledge gained through informal observation over a year's time of the day to day living of patients on seven tuberculosis wards. After a series of preliminary tests for clarity and "ratability" of items, the behavior scale was reduced to the present 64 items. This array of items was then re-examined and grouped on a rational basis into behaviors concerning six areas of hospital adaptation. These six areas were: (*a*) regulations to protect patient's health, (*b*) regulations to protect the health of others, (*c*) personnel, (*d*) other patients, (*e*) hospital facilities, (*f*) interaction with people outside hospital.

Behavior rating scoring. Here again, the experimenters made an a priori judgment as to whether a particular observable behavior was least or most adaptive to prolonged hospitalization. Four raters, two nurses and two aides, who had usually attended the patient for several months, rated him on each item. Each item comprised only two alternatives which took the generic form of behaving adaptively or not behaving adaptively. An item was scored in the following manner. If the four raters checked the adaptive alternative the item was scored four, if three of the four raters checked the adaptive alternative the item was scored three, if two raters checked the adaptive alternative the score was two, if one rater checked the adaptive alternative the score was one, if no rater checked the adaptive alternative the score was zero. The zero through five scores were then placed on a nine-point scale and assigned scores of 1.5, 3.0, 4.5, 6.0, and 7.5, respectively. The patient's score for any area was the sum of the items comprising that area.

Table 1

Comparison of Regular and Irregular Discharge Groups on Prehospital Life Adjustment Scales
(N = 58) *

Scale	Regular		Irregular			
	Mean	*SD*	Mean	*SD*	*t*	*p*†
Area of family background	126.53	13.81	127.31	15.81	—	—
Parents' economic status	24.21	3.72	24.37	3.57	—	—
Parents' education status	6.62	2.62	7.21	2.55	—	—
Parents' occupation status	15.85	3.61	16.42	2.61	—	—
Completeness of family life	29.36	6.42	29.89	7.25	—	—
Security of childhood	50.49	5.63	49.42	7.85	.53	<.50
Area of education						
Attainment	3.72	1.47	3.26	1.05	1.06	<.20
Grade school*	37.50	5.62	33.47	6.88	2.20	<.01
High school*	28.88	5.10	24.67	5.95	1.55	<.10
Area of social relations	107.24	11.36	101.18	12.11	1.83	<.05
Childhood	38.49	3.61	37.47	3.78	.98	<.30
Adolescence	31.60	5.14	29.20	4.82	1.75	<.05
Adulthood	37.15	5.14	34.51	7.26	1.43	<.10
Area of health	50.65	5.76	50.68	6.64	—	—
Childhood	21.72	2.88	22.58	2.61	—	—
Adolescence	13.90	4.21	14.20	6.17	—	—
Adulthood	15.03	2.88	13.90	3.38	1.22	<.20
Area of religion	16.62	2.85	15.00	3.30	1.84	<.05
Childhood	6.21	.84	6.00	1.08	.75	<.40
Adolescence	5.56	1.45	4.84	1.70	1.81	<.05
Adulthood	4.85	1.35	4.16	1.30	1.97	<.05
Area of drinking	25.36	3.13	23.89	4.54	1.28	<.20
Adolescence	6.92	.82	6.47	1.43	1.25	<.20
Army	6.08	1.14	5.68	1.33	1.08	<.20
Adulthood	5.92	1.68	5.42	2.05	.93	<.30
Last six months	6.44	1.46	6.32	1.98	.23	<.80
Area of occupation	52.44	9.69	46.06	9.90	2.32	<.01
Childhood	4.41	1.40	4.37	.94	.13	<.80
Adolescence	2.90	1.37	2.58	1.27	.89	<.30
Adulthood	45.13	8.29	39.11	9.96	2.28	<.01
Area of the military	52.08	7.34	48.47	6.11	2.00	<.01
Area of marriage						
Marital status	5.77	2.46	4.89	2.92	1.13	<.20
Family life*	41.53	6.20	41.86	4.23	—	—

* Grade school, high school, and family life subscales were computed with N's of 55, 22, and 48, respectively.
† p values are for one-tailed tests. No p value was ascertained for means representing reverse directionality.

Measurement of Situations External to the Hospital

From the unstructured interviews with ten patients and from discussions with staff members, 20 specific problem situations were isolated. These were then formulated into 20 items comprising four areas. These areas were: (*a*) patient's future employability, (*b*) pressure from others to continue treatment, (*c*) current economic status, (*d*) other family responsibilities, i.e., children, family illness, etc.

Scoring. The 20 items were scored in the same manner as the life adaptation items. All possible responses to each item were ranked on an a priori basis from the least to most adaptive. The score for each possible response to an item was its rank position, with the least adaptive response receiving a score of zero. The responses to each item were then rescored on a scale from zero to nine.

All scales used in the present study placed each response to every item on an a priori nine-point scale. Such a scoring system served several purposes. First, it preserved the least to most adaptive directionality. Secondly, each item contributed equally to the total score. Finally, all scale scores were independent of the empirical distribution, thereby simplifying replication of the study.

Results

The first hypothesis may be stated as follows: Irregular discharge patients when compared with regular discharge patients will have demonstrated poorer adjustment in areas of living prior to hospitalization for tuberculosis. Table 1 indicates that this hypothesis was supported at the 5 per cent level of confidence in the areas of: (*a*) social relations, (*b*) religion, (*c*) occupation, and (*d*) the military. The hypothesis was not supported in the areas of (*a*) family background, (*b*) health, (*c*) drinking, and (*d*) marriage. Viewed developmentally, Table 1 shows that the irregular group demonstrated significantly poorer adjustment than the regular group in grade school, in social relations and religion during adolescence, and in religion, occupation, and the military in adulthood.

The second hypothesis stated: Irregular discharge patients when compared with regular discharge patients will demonstrate poorer adjustment to situations within the hospital. Table 2 indicates that this hypothesis is not supported in terms of the expression of attitudes toward ward regulations, hospital personnel, other patients, and the facilities offered by the hospital. However, Table 3 indicates that the second hypothesis is supported when the ward behavior of the two groups is compared. In all six areas measured the behaviors are in the predicted direction. A further examination of Table 3 indicates that specifically the irregular discharge group demonstrated significantly poorer behavioral adaptation to restrictions of behaviors considered to be detrimental to their own health and to the health of others. Their relationships

Table 2

Comparison of Irregular and Regular Discharge Groups on Attitudinal Adaptation to Hospitalization ($N = 58$)

Scale	Regular		Irregular		t	p*
	Mean	*SD*	Mean	*SD*		
Attitude toward:						
ward regulations	60.51	9.79	58.16	8.32	.96	<.30
personnel	86.19	15.68	86.88	14.41	—	—
other patients	21.87	2.93	20.53	4.00	1.34	<.10
hospital facilities	69.00	9.36	66.42	10.13	.93	<.30

* p values are for one-tailed tests. No p value was ascertained for means representing reverse directionality.

Table 3

Comparison of Irregular and Regular Discharge Groups on Behavioral Adaptation to Hospitalization

Scale	Regular		Irregular		χ_c^2*	p†
	Above Median	Below Median	Above Median	Below Median		
Behavior concerning:						
regulations to protect patients' health	23	16	6	13	3.84	<.01
regulations to protect others' health	21	18	2	17	9.90	<.001
personnel	22	17	5	14	4.54	<.01
other patients	20	19	6	13	1.99	<.10
hospital facilities	19	20	5	14	2.57	<.10
people outside hospital	19	20	7	12	.70	<.30

* Median test utilized because of heterogeneity of variance.
† p values are for one-tailed tests.

Table 4

Comparison of Irregular and Regular Discharge Groups on the Adaptiveness of Situations External to the Hospital

Scale	Regular		Irregular			
	Above Median	Below Median	Above Median	Below Median	χ_o^{2}*	p†
Adaptation concerning:						
patients' employability	18	21	5	14	2.04	<.10
pressure from others to continue treatment	16	12	4	8	1.92	<.10
current economic status	14	14	6	6	—	—
other family responsibilities	14	14	6	6	—	—

* Median test utilized because of heterogeneity of variance.

† p values for one-tailed tests. No p value was ascertained for scales where frequency above and below the median was equal for both groups.

with ward personnel were also significantly less adaptive.

The third hypothesis stated: The type of situations external to the hospital are less conducive to continued hospital treatment for the irregular than for the regular discharge patients. Reference to Table 4 indicates that this hypothesis was not supported in any of the four selected classes of situations.

Discussion

When the results of this investigation are viewed collectively it would appear that hospitalization is but the most recent of a long series of life situations in which the irregular discharge patient has demonstrated his inability to make an adequate adjustment.

In an earlier study, the present authors were critical of the "personality determinism" type of approach typically used in experimental studies of the irregular discharge problem. At that time we argued, "It seems equally reasonable to assume that other factors, entirely independent of personality, may be just as important correlates of such behavior as the personality that the patient 'brings with him' into his particular hospital environment" (2, p. 66).

Insofar as extra-personality determinants were measured in the present study, however, they did not serve to differentiate the irregular and regular discharge groups. The results of the present study would suggest instead that the best predictor (or determinant) of response to hospitalization is past behavior in a variety of situations, with specific situational differences in the conditions of hospitalization playing a less important role than we formerly might have hypothesized.

Differences between the regular and irregular discharge groups in past life adjustment are striking. In grade school the irregular discharge group was significantly more truant, had more difficulty with peers and teachers. In adolescence they had fewer friends, fewer social interests, and quit attending church earlier. In the military, they were more frequently disciplined, they more often disliked officers and they more frequently failed to gain in rank. Occupationally, they changed jobs more frequently because they were "tired of the job" or were dismissed; they held more menial jobs, were less well-paid, and were more frequently unemployed. In most of the prehospital behavior areas measured the mean of the irregular group is in the direction of maladjustment, and in over half of the areas significantly so.

This maladjustive pattern, as reflected in the gross examination of past behavior, is magnified by direct observation of behavior in the hospital. Here the behavior of this group, as independently rated by four observers long before voluntary discharge against medical advice, was in the direction of maladjustment on every variable measured. The maladaptive nature of their behavior in the hospital was most evident in their disregard of regulations designed to protect their own health and the health of others, and in their friction with hospital personnel.

One further point about the nature of the irregular discharge problem should be mentioned. A small group of individuals, probably no more than 10 or 15 per cent of all people hospitalized for tuberculosis, appear to constitute the entire irregular discharge problem. This small group, through repeated irregular discharges, builds up the reported annual "irregular discharge rate" to almost 50 per cent. For example, the nineteen patients in the present sample have already contributed an average of four irregular discharges. Thus, a relatively small group of maladjusted individuals seem to provide the source for the technically accurate but somewhat misleading statement that almost half of all hospitalized tuberculous patients leave the hospital against medical advice.

Summary

This study examined some relationships of the tuberculous patient's (*a*) family of orientation, (*b*) prehospital life adjustment, (*c*) verbalized attitudes toward situations in the hospital, (*d*) behaviors in the hospital, and (*e*) current outside-the-hospital situations, to his success or lack of success in remaining in the hospital until his treatment was completed. Information in the above five areas was collected on 140 currently hospitalized patients. Within nine months after examination, 19 patients voluntarily left the hospital against medical advice and 39 completed their hospital treatment. Comparison of these two groups disclosed no differences in family of orientation, verbalized attitudes toward the hospital, or in current outside-the-hospital situations. Significant differences between the groups appeared, however, in prehospital life adjustment and in behavior within the hospital. It was concluded that hospitalization is but the most recent of a long series of life situations in which the irregular discharge patient has demonstrated his inability to make an adequate adjustment.

Received June 27, 1955.

References

1. Hand, T. J. Personality characteristics of a tuberculosis group. *Amer. J. phy. Med.*, 1952, **31**, 95–101.
2. Moran, L. J., Fairweather, G. W., Morton, R. B., & McGaughran, L. S. The use of demographic characteristics in predicting response to hospitalization for tuberculosis. *J. consult. Psychol.*, 1955, **19**, 65–70.
3. Phillips, L., & Cowitz, B. Social attainment and reactions to stress. *J. Pers.*, 1953, **22**, 270–283.
4. Rorabaugh, Mildred E., & Guthrie, G. The personality characteristics of tuberculous patients who leave the tuberculosis hospital against medical advice. *Amer. Rev. Tuberc.*, 1953, **67**, 432–440.
5. Speth, E. W. An investigation of emotional and social concomitants in the irregular discharge rates of hospitalized patients with tuberculosis. Unpublished doctoral dissertation, Univer. of Pittsburgh, 1953.
6. Tempel, C. W., *et al.* An analysis of hospital records of patients discharged from a large tuberculosis service. *U. S. Armed Forces Med. J.*, 1953, **4**, 1719–1733.
7. Thurston, J. R., & Calden, G. Intelligence factors in irregular discharge among tuberculosis patients. *J. consult. Psychol.*, 1954, **18**, 404.
8. Tollen, W. B. Irregular discharge: the problem of hospitalization of the tuberculous. *VA Pamphlet 10–27*, 1948.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

The Use of The W-B Picture Arrangement Subtest as a Projective Technique[1, 2]

Boris Breiger
Chicago State Hospital

The W-B Picture Arrangements (PA) was used as a projective technique in a cross-cultural study (1) comprising 30 U. S. Caucasians, 20 Nisei, and 10 German refugees. All *S*s were male, except two refugees. Variables controlled were Full-scale IQ, college status, age, urban-rural residence, and bilingualism. After completing the entire W-B, the "Flirt" and "Taxi" sequences were re-presented according to the *S*s' original arrangements. *S*s were asked to entitle the episodes and tell the story, card by card. A TAT-type inquiry followed.

Although no significant group differences were found on PA weighted scores or on raw scores of all sequences, a content analysis of "Flirt" and "Taxi" revealed marked differences. Categorization of "Flirt" and "Taxi" into projected motivational factors reveal chi-square values significant at the .01 and .05 levels respectively. Eighty per cent of the Caucasians projected a flippant story on "Flirt," entitling it "Pick up," while 65 per cent of the Nisei denied any romantic implication and entitled it "Chivalry." Nisei perceive the person in "Taxi" as being anxious in response to a member of his family "observing" him, while most Caucasians make no reference to any specific external source of anxiety. Significantly more Caucasians than Nisei (.01 level) projected abnormal sex behavior into "Taxi." Refugee stories followed U. S. Caucasian patterns.

Subsequent use of this technique with hospitalized patients aided in the diagnosis of well-defended persons who appeared intellectually intact upon standard W-B testing. Such persons frequently project stories expressing sex role confusion, prostitution fantasies, and paranoid ideation.

The present work suggests that an interpretation of the PA test based solely upon weighted scoring may be misleading. The "correct" arrangement may cover up perceptual distortions and other pathological processes. Failure to verify predicted relationships between W-B and Rorschach factors having a common rationale (2) may be due to overlooking the projective functioning of certain W-B stimuli. Awareness of these implications offers a lead for research seeking to integrate psychometric and projective test data.

Brief Report.
Received November 1, 1955.

References

1. Breiger, B. Cultural differences as indicated by an analysis of W-B subtests: with special reference to problems of clinical diagnosis and acculturation. Unpublished master's thesis, Univer. of Illinois, 1952.
2. Holzberg, J., & Belmont, L. The relationship between factors on the W-B and Rorschach having common psychological rationale. *J. consult. Psychol.*, 1952, **16**, 23–29.

---

[1] Presented at the 1953 meeting of the Midwestern Psychological Association, Chicago, Ill.

[2] An extended report of this study may be obtained without charge from Boris Breiger, Chicago State Hospital, Chicago 34, Ill., or for a fee from the American Documentation Institute. To obtain it from the latter source, order Document No. 4749 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Validity of the Marsh-Hilliard-Liechti MMPI Sexual Deviation Scale in a State Hospital Population

Roland M. Peek and Lowell H. Storms
Hastings, Minnesota, State Hospital

In a recent article in this Journal, Marsh, Hilliard, and Liechti (1) reported an MMPI scale of 100 items which served to pick up 88 per cent of two samples of sexual psychopaths at the expense of only 11 per cent false positives among UCLA students. However, a check on the utility of the scale for discriminating sex offenders from neurotics and psychotics revealed overlap so great as to suggest ineffectiveness, and the authors say that perhaps "some factor of personality integration or adjustment is being measured by the scale" (1, p. 58). This view is supported by their findings that high scorers on the scale obtained significantly higher *F* and lower *K* values. The present study was designed to investigate the effectiveness of this scale in a hospital setting by (*a*) evaluating its power to discriminate sex offenders from nonoffenders in small samples from the same hospital population, (*b*) comparing the scores of both groups with a nonhospitalized nonoffender group, and (*c*) comparing scores with other MMPI variables to test further the hypothesis that this scale is primarily a measure of gross personality integration. The present situation is less artificial than that of comparing college students with already referred sexual psychopaths, since the present experimental design more closely approximates the actual clinical situation and types of subjects with which the scale would be utilized.

Procedure

The Marsh-Hilliard-Liechti scale was scored on available MMPI's for three groups of subjects: (*a*) a group of 13 sex offenders comprising all those testable male patients in the Hastings (Minn.) State Hospital whose commitments were primarily for severe sexual misbehavior (four molesters, two rapists, two uncontrolled homosexuals, three exhibitionists, a fetishist, and a voyeur). Their mean age was 39.7 with a range of 16 to 62. Nine carried diagnoses of some form of psychosis, while four were personality or character disorders. Nine were single. (*b*) Thirty unselected male patients representing all those taking the MMPI consecutively during the period of the study and whose ages fell between 16 and 60, excluding the sex offenders, of course. Ages ranged from 19 to 52 with a mean of 36.8. In this group were 22 psychotics, three neurotics, and five personality and character disorders. Ten of the 30 were single, a significantly smaller proportion than in the above group ($p < .05$). (*c*) Thirty male psychiatric aides who took the MMPI as part of a routine battery administered soon after employment. Ages ranged from 20 to 56 with a mean of 31.3. Sixteen of the 29 whose marital status was known were single.

Mean score and overlap comparisons between the group of sex offenders and the patient and psychiatric aide groups were obtained, as well as comparisons with the original standardization groups (1). In the patient group, correlations were computed between the sexual deviation scale and each other MMPI scale.

Results

Mean scores on the sexual deviation scale were 43.31 ($SD = 7.88$) for the sex offenders,

Table 1

Discrimination of Sex Offender, State Hospital, and "Normal" Groups on the Sexual Deviation Scale, Using Original and Modified Cutting Scores
(Entries are percentages)

Cutting score	Sex offenders M-H-L (1) ($N = 338$)	Sex offenders Hastings State Hospital ($N = 13$)	Unselected State Hospital patients ($N = 30$)	Psychiatric aides ($N = 30$)	College students M-H-L (1) ($N = 317$)
>30	88	85	87	43	11
≦30	12	15	13	57	89
Total	100	100	100	100	100
>42	41*	77	33	7	1
≦42	59*	23	67	93	99
Total	100	100	100	100	100

* Interpolated from Marsh *et al.*, Table 2 (1).

38.77 ($SD = 6.65$) for the patients, and 29.70 ($SD = 8.61$) for the psychiatric aides.

Table 1 presents percentages of the Marsh *et al.* samples and the present groups falling above and below the original cutting score of 30 and the cutting score of 42 found to discriminate optimally between hospitalized sex offenders and other patients. The original cutting score, reported as picking up 88 per cent of California sexual psychopaths, similarly discriminates 85 per cent of hospitalized sex offenders in the present study. While this discrimination is attained at the expense of misclassifying only 11 per cent of the college students, 87 per cent of our patients (almost complete overlap) and almost half of the psychiatric aides are misclassified as sex deviates. The higher cutting score of 42 reduces the false positives among patients to 33 per cent and the false positives among college students and psychiatric aides to 1 per cent and 7 per cent respectively. However, this score, which achieves maximal differentiation in the hospital population, correctly identifying 77 per cent of the hospitalized sex offenders, selects only 41 per cent of the original group of sexual psychopaths.

Using the chi-square test, differentiations were significant at the 1% level between both sex offender groups and both "normal" groups. The discrimination between our sex offenders and patients was also significant at the 1% level, but only using the new cutting score of 42. All other comparisons were not significant, including three of the four comparisons of sex offenders with patients.

Correlations of the sexual deviation scale with MMPI scales computed for the 30 unselected patients are presented in Table 2. It can be seen that this scale correlates significantly with a number of MMPI scales, the strongest relationships being with the validity scales *F* and *K* (about .60, $p < .001$). Other reasonably high correlations (.50 to .53, $p < .01$) were found with *Mf*, *Ma*, and *Sc*. The

Table 2

Correlations of Sexual Deviation Scale with Other MMPI Scales in 30 State Hospital Patients

MMPI Scale	Correlation with Sexual Deviation Scale	Significance of correlation (t test)
L	−.30	ns
F	.60	.001
K	−.59	.001
Hs	.37	.05
D	.27	ns
Hy	.23	ns
Pd	.42	.05
Mf	.53	.01
Pa	.28	ns
Pt	.39	.05
Sc	.50	.01
Ma	.52	.01
Si	.20	ns

lower correlations with *Pd, Pt,* and *Hs* were significant at the 5% level.

Discussion

Our results confirm the incidental findings of Marsh *et al.* that their "sexual deviation" scale does not discriminate sex offenders from neurotics and psychotics, throwing more doubt upon their main conclusions than they indicated. The mean score of our patients was almost as high as the means of the sex offender groups and considerably above the reported cutting point; the mean of our "normal" group of psychiatric aides was almost at the cutting point they recommend. This cutting point fails completely to discriminate sex offenders from unselected state hospital patients, and even classifies two-fifths of the psychiatric aides with the committed sex offenders. The optimal cutting point for our groups provides better discrimination but the overlap remains unsatisfactorily large. Applying the base-rate analysis of Meehl and Rosen (2), the use of this scale to increase predictive success becomes still more questionable. Assuming a liberal 90 per cent discrimination of sex offenders as against 11 per cent false positives in a group to be delineated from them, with a base rate of 2 per cent for sex offenders in a given population (our rate is 1.6 per cent of the patients on the hospital rolls at the present time), 86 per cent of the persons classified as sex deviates by the scale will be false positives, actually not sex offenders.

Although computed on a small *N*, the correlations are consistent with the hypothesis that the "sexual deviation" scale is a measure of personality disintegration and general abnormality. The fact that the validity scales *F* and *K* correlate most strongly with the "sexual deviation" scale and in positive and negative directions respectively shows that higher scores on the sexual deviation scale are associated in this abnormal group with greater disorganization and lower defensiveness. This is interesting in view of the fact that item overlaps between this scale and the MMPI scales are very low (for example, three items overlap with *F;* seven overlap with *K*, but three of these are scored in the opposite direction).

Having confirmed the suspicion of Marsh *et al.* that their scale might be a measure of gross adjustment, our results indicate further that it is not a measure of sexual deviation in any broad application. A measure which differentiates a special group from another group may measure not only what the members of the special group have in common, e.g., sexual deviation, but may measure what that group has in common with any number of other groups, in this case general abnormality. It would be absurd, for instance, to call the Miller Analogies Test a Theoretical Physicist Scale because it discriminates theoretical physicists from high school graduates.

Summary

This study was performed to ascertain the validity of the Marsh, Hilliard, and Liechti Sexual Deviation Scale for the MMPI in groups of considerably different composition from their standardization samples and to check their incidental suggestion that the scale might be a measure of general personality integration.

Groups of 13 sex offenders, 30 psychiatric patients, and 30 psychiatric aides with available MMPI's were scored on the Sexual Deviation Scale, and correlations between that scale and the other MMPI scales were obtained for the group of patients. Comparisons were made among the present groups and between them and the original standardization groups using the above authors' cutting score as well as another which was found optimal for the present study.

The following results were obtained:

1. Neither cutting score differentiated effectively between patients and sex offenders; in fact, just as high a percentage of patients as of either group of sex offenders exceeded the cutting score recommended by Marsh *et al.*

2. The original cutting score failed to discriminate psychiatric aides from sex offenders satisfactorily, although the new cutting score was a little more successful.

3. Correlations of the scale with other MMPI scales support the hypothesis that the scale measures gross maladjustment or lack of personality integration. The strongest relationships are with the validity scales *F* and

K, and are positive and negative, respectively. High scorers on the "sexual deviation" scale tend to be more unstable, uncontrolled, confused, and prone to reality distortion.

It was concluded that the scale has little utility for the purpose for which it was designed and would fail to select sex offenders from many populations, since most people with high scores would have other than sexual problems. This scale seems to function more as a measure of general abnormality than of deviant sexual trends.

The inefficiency of such a scale in predicting behavior which is so rare in almost any population as are sex offenses was also discussed.

Received July 25, 1955.

References

1. Marsh, J. T., Hilliard, Jessamine, & Liechti, R. A sexual deviation scale for the MMPI. *J. consult. Psychol.*, 1955, **19**, 55–59.
2. Meehl, P. E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.*, 1955, **52**, 194–216.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Cognitive, Affective, and Psychopathological Correlates of the Taylor Manifest Anxiety Scale

Aron W. Siegman

Veterans Administration Hospital, Bronx, New York[1] *and Teachers College, Columbia University*

Though the Taylor Manifest Anxiety Scale (MAS) (21) is widely used as a measure of manifest anxiety, it has never been satisfactorily validated. Some attempts at validation have yielded positive results (4, 9), but others have yielded equivocal (10) or negative (2) findings. Moreover, recent research (11) suggests that a low score may not necessarily indicate lack of manifest anxiety, but rather *S*'s reluctance to admit it.

On the positive side are the many experiments which confirm predictions made on the assumption that the scale is a measure of manifest anxiety. In some (16) of these experiments it was found that much manifest anxiety, as indicated by a high Taylor score, enhances the performance on simple learning tasks but interferes with complex learning tasks. However, experimenters (7, 11) who obtained negative correlations between Taylor scores and scores on intelligence tests raised the question whether differences found in performance on learning tasks as a function of high and low Taylor scores might not really be a function of differences in intelligence. In most of these experiments, however, the intelligence test consisted of a limited number of tasks. Moreover, some of the tasks were actually learning tasks, and others had time limitations. In both cases a negative correlation is to be expected (16, 13, 19). In the present experiment Taylor scores were correlated with scores on the Wechsler Adult Intelligence Scale which includes a wide variety of cognitive tasks, timed as well as untimed, and only a few learning tasks.

Procedure

Several validation procedures were attempted. In the first, the Taylor MAS scores of 90 patients were correlated with their scores on a manifest anxiety rating scale (MARS). In the MARS, each patient was assigned by his psychiatrist or psychologist a score of 1, 2, or 3, corresponding to low, medium, and high, on nine criteria of manifest anxiety. These criteria are essentially the same as those described by Buss *et al.* (4). All the patients were in contact, and of at least normal intelligence. None was suspected of having cortical damage. In a further attempt to validate the Taylor MAS, the Taylor scores of patients who received a diagnosis of anxiety neurosis were compared with the Taylor scores of patients with other neurotic diagnoses but who were free from symptomatology of manifest anxiety. The Taylor scores of patients with other psychiatric diagnoses, particularly the psychopathic personalities, who are supposed to have very little manifest anxiety (1), were also of interest. In each case the diagnosis was agreed upon by the patient's psychiatrist and consulting psychologist, and all diagnostic "problem cases" were excluded. In yet another but more indirect attempt at validating the scale, the Taylor scores of all 90 patients were correlated with their scores on a self-esteem scale (20). Since a variety of dynamically oriented theorists (14, pp. 182, 191) maintain that anxiety is a function of threat to one's self-esteem, a negative correlation between the two scales was anticipated.

---

[1] From the Clinical Psychology Section of the Neuropsychiatric Service at the Veterans Administration Hospital, Bronx, N. Y. The author wishes to express his appreciation to Drs. H. L. Flowers, Chief of Neuropsychiatry and R. S. Morrow, Chief of Clinical Psychology for their sustained interest and support, and to Dr. Julia C. Hall, Assistant Chief Clinical Psychologist in charge of research, for her encouragement and help at every stage of this research.

To investigate the hypothesis that reluctance to reveal one's anxiety symptoms to others is a significant source of variance in performance on the Taylor MAS, the scale was administered under two different conditions. One group of 23 psychiatric but non-psychotic patients took the test as part of a diagnostic testing procedure, which they know has serious consequences. Another similar group of 33 patients took the scale in a group. They were told that since *E* was solely interested in the efficacy of the scale, and how people respond when they are free to be frank, they were not to indicate their names. Both groups were equated for age, IQ, and diagnosis. If reluctance to reveal one's anxiety symptoms depresses one's Taylor score, then the first group should obtain significantly lower scores than the second group.

Finally, in order to investigate the relationship of Taylor MAS scores with IQ, the Taylor scores of 35 medical and psychiatric patients, in whom brain damage and psychosis were definitely ruled out, were correlated with their WAIS IQ score. Their Taylor scores were also studied in relation to their scores on each of the eleven subtests, unadjusted as well as adjusted for IQ. Adjustment for IQ was achieved by subtracting *S*'s weighted score on a particular subtest from his mean weighted score on the remaining subtests.

Table 1

Mean Taylor MAS Scores in Relation to Psychopathology

Psychopathology	*N*	Mean	*SD*
Anxiety reaction	14	35.00	8.02
Other neuroses	18	22.89	12.04
Latent schizophrenia	17	22.53	12.17
Paranoid schizophrenia	15	26.27	11.44
Psychopathic personality	10	13.83	5.29

Results

The correlation between the patients' scores on the MARS and their Taylor scores was $r = .34$, which for 89 *df* is significant beyond the .01 confidence level.

Table 1 indicates that of the 90 patients, 14 received a diagnosis of anxiety neurosis, 18 received other neurotic diagnoses, including character neurosis, but were free of manifest anxiety symptomatology, 14 were schizophrenics, 17 were latent schizophrenics (18) and 10 received a diagnosis of psychopathic personality, which does not include the other character disorders (1). Of the remaining 16 patients, 5 had diagnoses which were not frequent enough to be grouped, 6 had neurotic diagnoses other than anxiety reaction but with manifest anxiety symptoms, and 5 were diagnostic "problem cases." Table 2 indicates that the patients with a diagnosis of anxiety reaction had significantly higher scores than

Table 2

t Values of the Differences Between Mean Taylor MAS Scores of the Various Psychopathological Groups

Psychopathological groups	*t*	*df*	*p*
Anxiety reaction vs. other neuroses	3.24	30	$<.01$
Anxiety reaction vs. latent schizophrenia	3.21	29	$<.01$
Anxiety reaction vs. paranoid schizophrenia	2.37	27	$<.05$
Psychopathic personality vs. anxiety reaction	6.97	22	$<.01$
Psychopathic personality vs. other neuroses	2.24	26	$<.05$
Psychopathic personality vs. latent schizophrenia	2.07	25	$<.05$
Psychopathic personality vs. paranoid schizophrenia	3.19	23	$<.01$
Paranoid schizophrenia vs. latent schizophrenia	.88	30	
Paranoid schizophrenia vs. other neuroses	.82	31	

Table 3

Mean Taylor MAS Scores in Relation to Test Identification and Test Anonymity

Group	*N*	Mean	*SD*	*t*
Test identification	23	27.30	10.65	
Test anonymity	33	31.64	7.96	1.82

any other group, and that the patients with a diagnosis of psychopathic personality had significantly lower scores than any other group.

The correlation between the patients' Taylor scores and their self-esteem scores was $r = -.72$, which for 89 *df* is highly significant.

Table 3 indicates that *S*s who took the Taylor MAS in a group and did not identify themselves failed to obtain significantly higher scores than *S*s who took the scale as part of a diagnostic testing procedure. This finding strongly suggests that reluctance to admit one's anxiety symptoms is not a significant variable.

Finally, there was no significant correlation between Taylor MAS scores and WAIS IQ scores ($r = -.19$, epsilon-square $= .16$). Furthermore, none of the correlations (*r*'s and etas) between the Taylor MAS and any single WAIS subtest, adjusted as well as unadjusted for IQ, was significantly different from zero.

Discussion

Though all three validating procedures yielded positive results, the correlation between the Taylor MAS and the MARS, the most direct criterion of manifest anxiety used in this study, is rather low. This finding is consistent with previous research which has had no difficulty in obtaining significant correlates of the Taylor MAS, but was much less successful in correlating the scale with direct criteria of manifest anxiety. In terms of a recent article by Cronbach and Meehl (5), these findings suggest that though criterion validation of the Taylor MAS yields questionable results, it does appear to have considerable construct validity.

There may be several reasons for the low correlation between the Taylor MAS and the MARS. One possible explanation is that the Taylor score represents the *S*'s evaluation of his own symptoms, and it is conceivable that two *S*s who are rated alike in manifest anxiety, one will be more sensitive to his own psychological experiences and symptoms than the other. This hypothesis is supported by the following facts: (*a*) previous investigators have reported higher correlations between the Taylor MAS and the *S*s' reports of their own manifest anxiety symptoms than between the Taylor MAS and the same but objectively observed symptoms (4), (*b*) the average MMPI profile of high Taylor scorers suggests that they are "introspective, quite sensitive to environmental press, and willing to admit to being easily disturbed. The low scorers, conversely, utilize both denial and repression fairly frequently and rarely introspect" (3, p. 435), and (*c*) several investigators (3, 6) have reported significant correlations between the Taylor MAS and the *Pt* scale of the MMPI, whose items refer to "unreasonable fears as well as overreaction to more reasonable stimuli" (8, p. 20). Another possible explanation for the low correlation between the Taylor scale and the MARS is that the Taylor scale is loaded with chronic somatic manifestations of anxiety (headaches, stomach troubles, etc.) and thus it is biased in the direction of a particular manifestation of anxiety. In order to test this hypothesis, the Taylor MAS was submitted to five experienced clinical psychologists who were asked to indicate which items reflected chronic somatic manifestations of anxiety and which reflected primarily psychic manifestations of anxiety. A chi-square test was then used to determine which items of the Taylor MAS differentiated at the .10 level or better between those *S*s who fell in the upper third and those *S*s who fell in the lower third of the MARS scores distribution. Whereas 12 items were judged as referring to chronic somatic manifestations of anxiety, only 1 such item (number 10) differentiated between the two groups at the .10 level (Table 4). Furthermore the correlation between the two sets of items with 59 *df* was only $r = .65$. These findings tend to support the hypothesis that the many items referring to chronic somatic symptoms are an undesired source of variance.

The absence of significant correlations be-

Table 4

Taylor MAS Items Which Differentiated at the .10 Confidence Level or Better Between a Group Which was Rated High and a Group Which was Rated Low in Manifest Anxiety

Item	Confidence level
1. I am no more nervous than most others.	.10
2. I work under a great deal of strain.	.01
3. I worry over money and business.	.02
4. I frequently notice my hand shakes when I try to do something.	.02
5. I worry quite a bit over possible troubles.	.01
6. I am often afraid that I am going to blush.	.01
7. I have nightmares every few nights.	.10
8. At times I lose sleep over worry.	.01
9. I often dream about things I don't like to tell other people.	.10
10. I have a great deal of stomach trouble.	.10
11. I am easily embarrassed.	.02
12. I feel anxious about something or someone almost all of the time.	.001
13. I am happy most of the time.	.05
14. It makes me nervous to have to wait.	.10
15. At times I have worried beyond reason about something that really did not matter.	.05
16. I have been afraid of things or people that I know could not hurt me.	.05
17. Sometimes I become so excited that I find it hard to get to sleep.	.10
18. I am the kind of person who takes things hard.	.10
19. I am a very nervous person.	.001
20. Life is often a strain for me.	.01
21. I am not at all confident of myself.	.10
22. I am very confident of myself.	.10

tween the Taylor MAS and any one of the eleven subtests of the WAIS raises an interesting problem, since many of these subtests are considered in clinical practice (15) as reliable indices of manifest anxiety. In light of the fact that there is substantial experimental evidence (12, 17) for this clinical judgment, the finding is a challenge to the validity of the scale. However, two possible explanations may be offered. One is that the Taylor MAS is a measure of *S*'s general level of manifest anxiety, but in no way indicates how *S* responded to a particular anxiety-stimulating task. Secondly, in clinical experience one seldom if ever relies on only one subtest, and it is conceivable that the Taylor scores are related to performance on a combination of anxiety-sensitive subtests. In order to test this hypothesis the subtests were divided into anxiety tests, which include arithmetic, digit span, digit symbol, block design, and object assembly (15, 12, 17) and nonanxiety tests which include information, comprehension, similarities, vocabulary, and picture completion.

Table 5 indicates that the group which obtained high Taylor scores (36 and higher, which represents the 75th percentile and above) also obtained significantly lower scores on the anxiety tests than the group which obtained low Taylor scores (17 and less; which represents the 25th percentile and below). This finding is consistent with good clinical practice not to rely on a single subtest.

Summary

A validational study of the Taylor MAS found that:

1. Taylor MAS scores correlated $r = .34$ with a manifest anxiety rating scale which for 89 *df* is significant beyond the .01 confidence level. It was hypothesized that the low correlation is due to the fact that the Taylor MAS contains too many items referring to chronic manifestations of anxiety. This hypothesis was supported by an item

Table 5

Mean Scores on Anxiety and Nonanxiety Subtests in Relation to Taylor MAS Scores

Group	*N*	Mean anxiety tests	*SD*	Mean nonanxiety tests	*SD*	*t*†
High Taylor scorers	10	10.00	1.50	11.60	2.20	2.86*
Low Taylor scorers	10	10.75	1.98	11.77	2.38	1.33

* Significant at the .02 confidence level.
† Fisher's *t* for correlated means.

analysis of the scale. Individual differences in self-awareness and sensitivity to one's symptoms seems to be another source of variance of the Taylor scale.

2. Patients with a diagnosis of anxiety reaction had significantly higher Taylor MAS scores than all other diagnostic groups, whereas patients with a diagnosis of psychopathic personality had significantly lower scores than all other diagnostic groups.

3. Taylor MAS scores correlated − .72 with scores on a self-esteem scale which for 59 *df* is significant beyond the .01 confidence level. It was suggested that the Taylor MAS has considerable more construct validity than criterion validity.

4. Patients who did not identify their Taylor MAS records did not obtain significantly higher scores than patients who did identify their records.

5. Though there was no significant correlation between Taylor MAS scores and WAIS IQ scores, high Taylor scorers did obtain significantly lower scores on the anxiety-sensitive subtests than on the nonanxiety-sensitive subtests, whereas no such difference was noted for the low Taylor scorers.

Received July 6, 1955.

References

1. American Psychiatric Association. *Diagnostic and statistical manual.* Washington, D. C.: American Psychiatric Association, 1952.
2. Bitterman, M. E., & Holtzman, W. H. Conditioning and extinction of the galvanic skin response as a function of anxiety. *J. abnorm. soc. Psychol.,* 1952, **47**, 615–623.
3. Brackbill, G., & Little, K. B. MMPI correlates of the Taylor scale of manifest anxiety. *J. consult. Psychol.,* 1954, **18**, 433–436.
4. Buss, A. H., Wiener, M., Durkee, A., & Baer, M. The measurement of anxiety in clinical situations. *J. consult. Psychol.,* 1955, **19**, 125–129.
5. Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.,* 1955, **52**, 281–303.
6. Eriksen, C. W., & Davids, A. The meaning and clinical validity of the Taylor anxiety scale and the hysteria-psychasthenia scales from the MMPI. *J. abnorm. soc. Psychol.,* 1955, **50**, 135–138.
7. Grice, G. R. Discrimination reaction time as a function of anxiety and intelligence. *J. abnorm. soc. Psychol.,* 1955, **50**, 71–74.
8. Hathaway, S. R., & McKinley, J. C. *Minnesota Multiphasic Personality Inventory.* New York: Psychological Corporation, 1943.
9. Hoyt, D. P., & Magoon, T. M. A validation study of the Taylor Manifest Anxiety Scale. *J. clin. Psychol.,* 1954, **10**, 357–361.
10. Kendall, E. The validity of Taylor's Manifest Anxiety Scale. *J. consult. Psychol.,* 1954, **18**, 429–432.
11. Kerrick, Y. S. Some correlates of the Taylor Manifest Anxiety Scale. *J. abnorm. soc. Psychol.,* 1955, **50**, 75–77.
12. Mandler, G., & Sarason, S. B. A study of anxiety and learning. *J. abnorm. soc. Psychol.,* 1952, **47**, 166–173.
13. Matarazzo, J. D., Ulett, G. A., Guze, S. B., & Saslow, G. The relationship between anxiety level and several measures of intelligence. *J. consult. Psychol.,* 1954, **18**, 201–205.
14. May, R. *The meaning of anxiety.* New York: Ronald Press, 1950.
15. Mayman, M., Schaffer, R., & Rapaport, D. Interpretation of the Wechsler-Bellevue Intelligence Scale in personality appraisal. In H. H. Anderson & Gladys L. Anderson (Eds.), *An introduction to projective techniques.* New York: Prentice-Hall, 1951.
16. Mayzner, M. S. A critical review of recent experimental studies relating anxiety to learning. *Psychol. Newsltr.,* 1954, **5**, 117–137.
17. Moldawsky, S., & Moldawsky, Patricia C. Digit-span as an anxiety indicator. *J. consult. Psychol.,* 1952, **16**, 115–118.
18. Shapiro, D. Special problems of testing borderline psychotics. *J. proj. Tech.,* 1954, **18**, 387–395.
19. Siegman, A. W. The effect of manifest anxiety on a concept formation task, a nondirected learning task and on timed and untimed intelligence tests. *J. consult. Psychol.,* in press.
20. Siegman, A. W. Some factors associated with visual threshold of taboo words. *J. clin. Psychol.,* in press.
21. Taylor, Janet A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.,* 1953, **48**, 285–290.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Reliability of Case History Ratings and Intellectual Ability of Graduate Raters

A. W. Bendig
University of Pittsburgh

Two studies have related educational level of raters to several measures of the reliability of ratings of clinical case histories (1, 2). Three rating measures were used: *rater reliability*, which is the average of the $N(N-1)/2$ intercorrelations among the N raters in a group; *psychiatric reliability*, the average of the N correlations of each single rater's judgments with the pooled judgments of psychiatrists of the same clinical material; and *rater bias*, which is a measure of the variability among the N raters in the mean rating each rater assigns all of the case histories he judges. Bendig and Sprague (2) found undergraduate raters to be lower in rater reliability and higher in rater bias than graduate students in psychology and practicing clinical psychologists. Bendig (1) used a different set of case histories and reported no significant differences between groups of undergraduate and graduate *Ss* in rater and psychiatric reliability, but again the undergraduate raters demonstrated a significantly greater variability in bias measures (1, p. 129).

These differences in rater bias can be assumed to be the result of the educational differences between the *Ss*, but whether less bias is due to greater training in psychology, higher intellectual level, or increased interest in clinical work with patients among graduate raters is unknown.

Procedure

Case histories. The rated stimuli were the 10 abstracted clinical case histories used in a previous study (1). The cases varied in length from one to two pages of single-spaced typing and had been abstracted from complete case histories in clinic files. The clinical subjects of the cases were all white males and varied in adjustment level from relatively normal to hospitalized schizophrenic. The abstracted cases were mimeographed as a booklet with a face sheet containing rater instructions, the rating scale to be used, and spaces for recording *S*'s judgments. The scale was a seven-category scale with the center and end categories verbally anchored by the descriptive phrases "slightly maladjusted," "moderately maladjusted," and "extremely maladjusted." The intermediate scale categories were left unanchored as was done in previous use of this scale (1, 2).

Subjects. The raters were 40 graduate students in psychology, ranging in graduate training from first- to fourth-year students. Of this group 25 had expressed their major area of professional interest as clinical or counseling psychology, with the remaining 15 having their major interest in other areas, i.e., experimental, comparative, social, psychometrics, etc. Most of the *Ss* ($N = 27$) had taken the three-test graduate school entrance battery described in detail by Jenson (5) and these scores were available from the files of the University Testing Service. The case history booklets and rating scales were distributed to the *Ss* in their graduate classes and the completed ratings collected at a subsequent class period.

Rater variables. The 27 *Ss* who had taken the graduate test battery were dichotomized on the basis of their raw scores on four intellectual variables: the Miller Analogies test, a Mathematical Aptitude test, a Reading Comprehension test, and a measure of Research Aptitude. This last variable had been

Table 1

Reliability of Case History Ratings of Graduate Psychology Students Dichotomized on Six Rater Variables

			Rater reliability		Rater bias		Psychiatric reliability	
Rater variable	Rater group	N	r	t	r	t	r	t
Miller Analogies	High	13	.81**	1.38	.69**	2.77**	.86**	.69
	Low	14	.86**		.44*		.89**	
Mathematical aptitude	High	14	.82**	1.23	.54*	.78	.87**	.53
	Low	13	.86**		.61**		.88**	
Reading comprehension	High	14	.84**	.31	.64**	.94	.87**	.30
	Low	13	.85**		.55*		.88**	
Research aptitude	High	15	.85**	.62	.59**	1.60	.89**	.86
	Low	12	.82**		.43		.86**	
Graduate credits	High	20	.87**	1.69	.73**	3.62**	.89**	1.15
	Low	20	.81**		.49*		.86**	
Clinical interest	High	25	.85**	.85	.63**	.01	.88**	.38
	Low	15	.82**		.63**		.87**	

* Significant at the .05 level.
** Significant at the .01 level.

developed by Hyman (4) through factor analytic procedures as a predictor of combined criteria of research ability and consisted of a linear combination of subscales from the Reading Comprehension test. The total group of 40 *S*s was also dichotomized on the basis of the total number of graduate credits earned in psychology and on the basis of their expressed clinical interest.

Results

The *S*s were dichotomized separately on each of the six rater variables and three reliability measures computed for each pair of groups. Two of these measures, rater reliability and rater bias, were obtained by analyses of variance of the ratings (1, 2). Rater reliability is the average intercorrelation among the raters in the ratings assigned to the stimuli, while rater bias is a measure of the individual differences among the raters within a group to consistently overrate or underrate all of the cases. The third measure, psychiatric reliability, was obtained by correlating each *S*'s ratings with the pooled judgments of six psychiatrists who had also evaluated the 10 cases (1). The average correlation of each group of *S*s with the psychiatrists was obtained by finding the mean of the transformed (r-to-z) correlations within that group (3, pp. 133–134).

The pairs of rater reliability and rater bias measures for each dichotomized group of raters were tested for significant differences by the procedure given by Edwards (3, p. 136), using as the degrees of freedom for each transformed coefficient one less than the number of degrees of freedom of the error term in the above noted analyses of variance. Pairs of psychiatric reliability coefficients were tested by similar procedures, using as the degrees of freedom for each averaged correlation the sum of the df for each S ($df = 7$ for one S). The results of these comparisons can be found in Table 1.

None of the dichotomized group differences in rater or psychiatric reliability was significant at the .05 level of confidence and only one of these twelve comparisons, that between *S*s high and low in the number of earned graduate credits, was significant at the .10 level ($t = 1.69$). Consequently it seems reasonable to conclude that intellectual and interest differences between these graduate *S*s had no effect upon their reliability in judging clinical case histories. However, two of the six comparisons on rater bias were significant at the .01 level. The *S*s with high scores on the Miller Analogies test showed significantly more rater bias ($t = 2.77$) and *S*s having earned more graduate credits also showed more bias ($t = 3.62$). The significant differences between *S*s dichotomized on the basis of these two rater variables cannot be explained by a relation between Miller Analogies test scores and number of graduate credits since the product-moment correlation between these two variables for 27 *S*s was only $-.05$.

Discussion

Apparently intellectual, training, and interest differences among graduate students have no effect upon the reliability of their ratings of clinical case histories, but do affect measures of rater bias. One *S* may rank order the 10 cases about the same as another *S*, but the first rater spreads the cases along the low or adjusted end of the rating scale continuum while the second rater distributes the same case histories along the high or maladjusted portion of the scale. Raters with high Miller Analogies scores tended to vary more among themselves in the central tendency of their ratings than do *S*s with low scores. Individual differences in the portion of the scale used for rating are also greater for more educationally advanced graduate students. This greater rater bias is possibly a function of the divergent connotations of the global concepts of "adjustment" among more intellectually able and more academically experienced graduate *S*s. For one rater the reported case behavior has to be quite deviant before he is willing to judge it as maladjusted, while the same behavior is judged much more stringently by another rater. Although these two rater variables, one intellectual and one educational, are related to the *absolute* adjustment level ratings given by graduate *S*s to case histories, they are not related to the *relative* ratings assigned by the raters.

Summary

Graduate students in psychology ($N = 40$) rated ten abstracted clinical case histories for global adjustment level using a seven-point rating scale. The raters were then dichotomized on the basis of their scores on graduate school entrance tests, the number of earned graduate credits in psychology, and their expressed interest in clinical psychology. Measures of rater reliability and bias were computed for each of the dichotomized groups and the measures of reliability and bias compared by *t* tests. None of the intellectual, educational, or interest variables were related to rater reliability, but raters high on the Miller Analogies Test and with a greater number of earned graduate credits showed significantly (.01 level) greater rater bias.

Received August 8, 1955.

References

1. Bendig, A. W. Rater experience and case history judgments of adjustment. *J. clin. Psychol.,* 1955, **11**, 127–132.
2. Bendig, A. W., & Sprague, J. L. Rater experience and the reliability of case history ratings of adjustment. *J. consult. Psychol.,* 1954, **18**, 207–211.
3. Edwards, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950.
4. Hyman, S. Development of criteria of research competence in psychology and their prediction from certain intellectual and achievement measures. Unpublished doctor's dissertation, Univer. of Pittsburgh, 1954.
5. Jenson, R. E. Predicting scholastic achievement of first-year graduate students. *Educ. psychol. Measmt,* 1953, **13**, 322–329.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Body Cathexis as a Factor in Somatic Complaints[1,2]

Laverne C. Johnson
Washington University School of Medicine and Research Laboratories of Malcolm Bliss Hospital

In recent work Secord (7) and Secord and Jourard (8) have focused attention upon attitude toward the body as an important variable in any comprehensive theory of personality development. They, like Murphy (6), conceive of the body as an anchorage point for the more inclusive concept of the self and by use of objective measures of body cathexis (*BC*) and self cathexis (*SC*) have found a significant relationship between the two. Body cathexis was defined by Secord and Jourard as the degree of feeling of satisfaction or dissatisfaction with the various parts or processes of the body. This definition will be used in this paper. In addition to cross validation of the relationship of body and self cathexis, this paper: (*a*) presents information concerning the stability of attitude toward the body and self, and (*b*) investigates the relationship of body cathexis to somatic complaints.

Even though Freud related hypochondriasis to an excessive attraction of interest onto the sensations and functions of bodily organs, few studies have been reported relating attitude toward the body to somatic complaints. Levy (3) investigated body interest and hypochondriasis in 20 children, but he was more interested in the types of body concern and causes of excessive body interest than he was in comparing these 20 children to a control group.

Secord and Jourard (8) utilized a homonym test composed of words having meanings pertaining to the body and also meanings not related to the body to study this question. The subject responded with the first word that came to him after each oral presentation of the homonym. A significant relationship between the number of responses having bodily meanings and *BC* was found. These results, Secord and Jourard believe, confirmed their hypothesis "that low *BC* is associated with anxiety in the form of undue autistic concern with pain, disease, or bodily injury" (8, p. 343). While these results indicate a relationship between *BC* and bodily concern when measured by a word-association technique, the findings do not answer the question whether *BC* will be related to the actual number of somatic symptoms experienced and reported. Will dissatisfaction with and, possibly, concern over body appearance and body functioning be related to the number of somatic symptoms reported on a health questionnaire? The question is investigated in the present study.

Subjects

Fifty-two males and ninety-five females with mean ages of 20.2 and 20.3 comprise the sample studied. As part of a larger research project,[3] the fifty-two males were drawn from the freshman class of a local Protestant seminary. These students were selected from the total class, $N = 183$, solely on the basis of their electroencephalographic records. None of the EEG records of the subjects chosen would be classified as abnormal. The ninety-five females were student nurses receiving their psychiatric training at Malcolm Bliss Hospital.

---

[1] This study was supported in part by a research grant from the National Institute of Mental Health of the National Institutes of Health, United States Public Health Service.

[2] Part of these data was gathered under Air Force Contract Number 18(600)–927 with the Department of Clinical Psychology, United States Air Force School of Aviation Medicine.

[3] Air Force Contract Number 18(600)–927, The Psychiatric Screening of Flying Personnel.

Procedures

Each subject was administered the *BC* and *SC* scales and the Cornell Medical Index Health Questionnaire. The male subjects were administered the tests as part of the more extensive battery when they came to our hospital laboratories.

The Tests

The *BC* and *SC* scales were very similar to those used by Secord and Jourard (8). Instructions for the scales were changed to reduce the tendency to response sets. The following instructions appeared on the cover page:

> On the following pages are listed a number of things characteristic of yourself or related to you. You are asked to indicate which things you worry about and would like to change if it were possible, and which things you have no feelings about one way or the other.
>
> Consider each item listed below and encircle the number which best represents your feelings about yourself now according to the following scale: *1.* Strongly dislike and wish change could somehow be made; *2.* Don't like, but can put up with; *3.* Have no particular feelings one way or the other; *4.* Definitely like, am pleased with; *5.* Consider myself particularly and unusually fortunate. So that you will be able to judge each item carefully in terms of the above five statements, the scale will be at the top of each page. You may refer back to the scale as often as necessary to make your judgment of how you feel. Judge each item carefully. Do not use the same number for each item.

The items and scoring procedure were essentially those used by Secord and Jourard. Organs pertaining to sexual functions were not excluded as was done by Secord and Jourard (8).

These instructions were very successful in reducing response sets. The criteria used in the original study to measure response set were as follows: (*a*) a frequency ≥ 32 in category 4; (*b*) a frequency ≥ 28 in category 5; and (*c*) a frequency ≥ 24 in category 5 accompanied by less than two responses in categories 1 and 2 combined. Applying these criteria to the present sample, only 12, 8 men and 4 women, of the 147 subjects would be suspected of response sets. When compared to the 37 of the 126 subjects eliminated by Secord and Jourard, response set does not seem to be a problem in this sample using these instructions.

The Cornell Medical Index Health Questionnaire consists of a series of questions which the subject is to answer yes or no. As the name implies, the items are concerned with symptoms and diseases past and present. The particular form used in this study is that used by the Air Force (1). With the exception of containing fewer questions, 171 in contrast to 195, the test used here is identical to the published Cornell Medical Index Health Questionnaire. Both men and women took the same form. The score on the Health Questionnaire consisted of the number of symptoms or diseases reported as indicated by the sum of the yes responses.

Results

The three scores obtained from the *BC-SC* scales were: total *BC*, total *SC* and a score called an "anxiety indicator" by Secord and Jourard. The first two scores were a total of the ratings for each item on a 5-point scale for each individual divided by 50 and 52 respectively to get the average *BC* and *SC* rating. A low score on the *BC* or *SC* scale indicates that the individual has a low attitude toward his body or self and wishes change could be made. A high *BC* or *SC* score on the other hand indicates the individual is pleased with his body or self. A low rating then is indicative of negative cathexis toward the body or self and a high score is indicative of positive cathexis. The "anxiety indicator" score was obtained by summing the ratings for each male on the 13 most negatively cathected items by the males and the 14 most negatively cathected items for the females. The sums were then divided by 13 and 14 to get an average rating.

Items most negatively cathected by the males were: strength, body build, muscles, sex activities, weight, waist, energy level, width of shoulders, hips, teeth, facial complexion, posture and nose. The most negatively cathected items for the females were: hair, facial complexion, nose, waist, body build, appearance, hips, legs, teeth, feet, knees, posture, face, weight. These items are almost identical to those reported by Secord and Jourard, and as in Levy's (3) study the items most negatively cathected indicate concern over body appearance rather than body functioning. Also, as reported by Levy, women are primarily concerned with appearance while men in addition to appearance are concerned with physical prowess.

As Taylor Manifest Anxiety scores were available as part of the total test battery, it was possible to obtain the correlation between the Taylor scale scores and the most negatively cathected items, the "anxiety indicator" score. Results indicate an inverse relationship between "anxiety indicator" scores and manifest anxiety. The correlations were − .40 for men and − .53 for women. Low attitude toward body appearance is significantly associated, beyond the .01 level, with manifest anxiety. The difference between the correlations for men and women was not significant.

Reliability and Relations of BC *and* SC

The split-half reliability reported by Secord and Jourard for the *BC* and *SC* scales indicated the internal consistency of the scale but gave no indication as to the stability of attitude toward the body and self over a period of time. Retesting of the 52 male subjects after a 6 to 8 week interval made possible the obtaining of the test-retest coefficients of reliability. These were .72 for *BC*, .74 for *SC*, and .76 for the "anxiety indicator" scale. These figures indicate acceptable test-retest reliability for the three measures.

Table 1
Means, Standard Deviations, and Comparison of Difference Between Means for Sexes
($N=52$ Males; 95 Females)

Test	Mean	*SD*	*CR*
Body cathexis			
Males	3.44	.50	2.13*
Females	3.27	.38	
Self cathexis			
Males	3.39	.53	NS
Females	3.38	.37	
Anxiety indicator			
Males	3.19	.67	3.06**
Females	2.86	.58	
Cornell Medical Index			
Males	88.02†	44.36	3.53**
Females	112.74†	32.28	

* Significant beyond .05 level.
** Significant beyond .01 level.
† Log transformed scores.

Table 2
Intercorrelations Between *BC* Scores and Cornell Medical Index

Test	Body cathexis	Anxiety indicator
Cornell Medical Index		
Males	−.33*	−.42**
Females	−.40**	−.54**

* Significant beyond .05 level.
** Significant beyond .01 level.

The relationship of *BC* to *SC* as indicated by Pearson product-moment correlations was .66 for male and .79 for females. These correlations support the findings by Secord and Jourard that attitude toward the body is related to attitude toward the self.

Sex Differences

The means, standard deviations, and comparison of difference between means for the two sexes are listed in Table 1. As the distribution of scores on the Cornell Medical Index was skewed with the variability being almost equal to the mean, these scores were converted to log scores and these transformed scores were used for all statistical analysis (5).

The student nurses' attitude toward their bodies was significantly lower than that of the seminary students on the total *BC* scale and on the "anxiety indicator" scale. Difference between the sexes on the latter which consists of items pertaining primarily to body appearance was particularly striking. When cutting scores of 2.8 or below and 4.0 or above are used for dividing the two groups, 61 of the women, 64%, in contrast to 22, 42%, of the men have scores below 2.8. Only 2 of the women, 2%, have scores of 4.0 or above while 7, 13%, of the men have scores in this range. Thus 64% of the women, in contrast to 42% of the men, state they definitely dislike the appearance of their body and only 2%, in contrast to 13%, state they are pleased with their appearance. Significantly more symptoms were admitted to by the student nurses on the Cornell Medical

Table 3

Intercorrelations of Somatic Items on Cornell Medical Index and *BC*

Test	Body cathexis	Anxiety indicator
Somatic items		
Males	−.24	−.31*
Females	−.30**	−.44**

* Significant beyond .05 level.
** Significant beyond .01 level.

Index, but there was no difference between the sexes on the *SC* scale.

Relationship of BC *to Somatic Symptoms*

The correlations between *BC* and the Cornell Medical Index for both sexes are listed in Table 2.

All correlations are significant beyond .05 level, and all but one are significant beyond the .01 level indicating in this sample a moderate inverse relationship between attitude toward the body and number of symptoms reported. While the correlations for women are higher than those for men the differences are not significant. For both sexes there is a higher relationship between the "anxiety indicator" scores and number of symptoms reported than when the total *BC* score is used. This difference is significant for women at the .05 level when checked by z transformation using the technique for difference between correlated z's (4, p. 148). The difference for the men is not significant.

As the Cornell Medical Index is composed of items pertaining to both somatic and psychological symptoms the question arose whether the intercorrelations were a result of psychological complaints rather than somatic ones. To check this the Cornell Medical Index was rescored for somatic symptoms only and these scores were correlated with *BC*. These results are summarized in Table 3. While the correlations between *BC* and only somatic symptoms are lower, all are significant with the exception of total *BC* and somatic items for the men. This reaches the .09 level.

Discussion

This study cross validates Secord's and Jourard's finding that attitude toward the body is a significant factor in attitude toward the self. Zachry (10) believes that one of the most important tasks of adolescence is the acceptance of the body "as a symbol of self." These results suggest that the attitude toward one's body is associated with acceptance of self, not only in adolescence, but also in young adults, and perhaps at all ages.

Results indicating a moderate relationship between attitude toward the body and somatic complaints renews interest in Freud's theory of hypochondriasis. In his chapter on narcissism Freud states: "The hypochondriac withdraws both interest and libido—the latter especially markedly—from the objects of the outer world and concentrates both upon the organ which engaged his attention" (2, p. 40). Freud believed that this withdrawal from external objects and excessive attraction of interest onto the sensations and functions of body organs was due to a damming back of the libido into the internal organs. This paper was not designed specifically to test Freudian theory, and the subjects have not been clinically diagnosed as hypochondriac, but the results support the hypothesis that attitude toward the body is associated with somatic complaints.

The significantly lower *BC* scores for women suggest the attractive hypothesis that because of the social importance of the female body women are more critical of their bodies and are likely to be more concerned over body appearance and functioning than are men. Jourard and Secord (9) investigated *BC* and the ideal female figure and found that women's satisfaction with aspects of their bodies varies with the magnitude of the deviation between measured size and what they consider ideal size. They also found that many women feel attainment of "ideal" proportions is difficult, if not impossible. It certainly seems plausible then that, insofar as this "ideal" is internalized by women, failure to attain it would result in low body cathexis. The small percentage of both men and women who stated they were pleased with the appearance of their bodies indicates how difficult it is to measure up to the "ideal."

However, certain factors suggest caution in accepting the above view completely. Our results are not compatible with Secord and Jourard (8) with regard to sex differences. They found no difference between male and female on either the body cathexis scale or homonym test of body concern.

When the *BC-SC* scores of this study were compared with those of Secord and Jourard, no difference was found on the *SC* scores for either sex or for men on the *BC* scale. However, a significant difference, at the .008 probability level, was found between the student nurses and college women on total *BC*. The difference between the two female groups was not significant on items pertaining primarily to appearance, "anxiety indicator" scale. Three factors may be related to these differences between the sexes and between the women of the two studies: age, intellectual level, and training.

Age was almost identical for the sexes in this study and is similar to that of college sophomores and thus safely can be ruled out as an important variable contributing to the differences. Intellectual level cannot be as safely excluded but as the mean IQ of the student nurses was 112, they appear to be not too dissimilar, intellectually, from students in many liberal arts colleges. Perhaps the most significant difference between the sexes, and the women of this study and those of Secord and Jourard's, is the training and duties of the student nurse. Levy (3) noted that one factor causing excessive body concern and interest in his sample of children was exposure to illness or being exposed to frequent discussions of physical ills. It is a common phenomenon for medical students to be stricken with various ailments and become concerned with the functioning of their bodies when they begin their clinical studies. Perhaps a similar phenomenon is present in the student nurses and their low scores on total *BC* and the larger number of symptoms reported is due in part to their exposure to and interest in illness.

Summary

The purposes of this study were (*a*) to cross validate the findings of Secord and Jourard that attitude toward the body is related to attitude toward the self; (*b*) ascertain the stability of *BC* and *SC;* and (*c*) to investigate the relationship of *BC* to somatic complaints. Fifty-two male seminary students and ninety-five student nurses were administered the *BC* and *SC* scales and the Cornell Medical Index. A significant relationship was found between *BC* and *SC*, cross validating Secord's and Jourard's findings. Attitude toward the body and toward the self was found to be stable over a period of time. A moderate inverse relationship was found between attitude toward the body and number of somatic symptoms reported. Women had lower *BC* scores than men and reported a larger number of symptoms. The social importance of the female body and the training and interest of the student nurses in illness are offered as possible explanations.

Received July 29, 1955.

References

1. Barry, J. R., & Raynor, G. H. Psychiatric screening of flying personnel. Research on the Cornell Index. *USAF Sch. Aviat. Med. Proj. Rep.*, 1953, Proj. No. 21-0202-007 (Rep. No. 2).
2. Freud, S. *Collected Papers.* Vol. IV. London: Hogarth Press, 1953.
3. Levy, D. M. Body interest in children and hypochondriasis. *Amer. J. Psychiat*, 1932, **12**, 296–315.
4. McNemar, Q. *Psychological statistics.* New York: Wiley, 1955.
5. Mueller, C. G. Numerical transformations in the analysis of experimental data. *Psychol. Bull.*, 1949, **46**, 198–223.
6. Murphy, G. *Personality.* New York: Harper, 1947.
7. Secord, P. Objectification of word-association procedures by use of homonyms; A measure of body cathexis. *J. Pers.*, 1953, **21**, 479–495.
8. Secord, P. F., & Jourard, S. M. The appraisal of body-cathexis: body-cathexis and the self. *J. consult. Psychol.*, 1953, **17**, 343–347.
9. Jourard, S. M., & Secord, P. F. Body-cathexis and the ideal female figure. *J. abnorm. soc. Psychol.*, 1955, **50**, 243–246.
10. Zachry, C. B., & Lighty, M. *Emotion and conduct in adolescence.* New York: Appleton-Century-Crofts, 1940.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Stability of the WISC and Binet Tests

Ila H. Gehman
State Hospital at Butner, N. C.[1]
and Rudolph P. Matyas
Pennsylvania State University

The present study explored the stability of the Wechsler Intelligence Scale for Children and the Revised Stanford-Binet Scale, Form L, as measures of intelligence over a four-year time interval. No information has been made generally available to show the long-term reliability of the WISC. The 1937 edition of the Binet has not been extensively investigated with general populations in this respect, but Matyas' (3) survey of the literature suggested that an r of approximately .73 was typical of test-retest correlations when the intervals between tests were of such duration as to minimize practice effects.

It was hypothesized that intelligence quotients derived from the WISC would be found to be as constant upon retest after a period of four years as would those derived from the Binet. The examiners were also interested in checking upon the relative efficiency of administration of the two scales, and kept account of the time required to give each test.

Subjects and Procedures

Sixty pupils in the ninth grade of a junior high school in central Pennsylvania participated in this study. They were selected on the basis that their scores on the WISC and Stanford-Binet, Form L, were available from research by Clarke (1) completed in May, 1950. The sample in Clarke's study included 85 children who comprised the entire population of three fifth grade classes in the school district. Their mean age was 11 years, 1 month. At the time of the present investigation, 60 of the original population were available for examination and their mean age was 15 years, 2 months. Twenty-nine of the 60 pupils were boys and 31 were girls.

The WISC and Binet were administered during the spring of 1954 by three graduate students who had completed courses in individual testing taught by the senior author. Three other examiners who had also been instructed by this same person examined these children when they were in fifth grade in 1950. Testing conditions in the two investigations were similar and it was generally possible to alternate the administration of the two tests. Data for the 60 pupils from the original sample of 85 in the fifth grade were recomputed in order to provide a basis for direct comparisons of identical subjects for the two age and grade levels.

Results and Discussion

Table 1 describes the composition of the group with respect to IQ's obtained on first and second examinations. The means indicate a population with average general intelligence.

In the fifth grade none of the means was found to be significantly different from any of the others at the 5% level of confidence, but in the ninth grade the mean WISC Performance IQ was significantly greater than the WISC Verbal IQ. Sixty-seven per cent of the ninth grade pupils had higher performance than verbal IQ's. The difference of 6.51 IQ points is almost identical with a finding from Delattre and Cole (2) working with children of ages 10–5 to 15–7 who tended to be above average in intelligence. These discrepancies suggest a possibility that the verbal and performance scales are not of equal diffi-

---

[1] Formerly Associate Professor of Psychology, Pennsylvania State University.

Table 1

Intelligence Quotients of Sixty Children Retested After Four Years

Test	Grade 5, 1950 Mean Age 11-1 Mean IQ	 *SD*	Grade 9, 1954 Mean Age 15-2 Mean IQ	 *SD*	Difference	*t*
Binet L	96.17	14.29	98.25	14.38	2.08	.79
WISC Verbal	96.90	10.84	96.72	11.14	−.18	.06
WISC Performance	99.87	13.19	103.23	14.44	3.36	1.40
WISC Full Scale	98.13	11.20	99.82	12.79	1.69	.76

Table 2

Correlation Between IQ's from Tests Given in Fifth and Ninth Grades

Test	*r*
Binet L	.78
WISC Verbal	.77
WISC Performance	.74
WISC Full Scale	.77

culty at the older age levels. However, among other speculations concerning the cause of the discrepancies it is possible that the performance test is more sensitive to "coaching," when intensive testing is done in one school system, than is true of the verbal scale.

While there was a slight tendency for the IQ's to rise from grade five to grade nine (except in the WISC Verbal), none of the *t* tests reached significance at the 5% level. Thus the mean IQ's were stable for both the Binet and all scales of the WISC over the four-year period.

The relatively small sigmas would tend to reduce the correlations between the first and second testings. This restricted variability is largely a function of the selection of students from specific grade levels. Even so, the correlations were all found to be high and significant as is shown in Table 2. The stability coefficients were markedly similar with none of the correlations differing from any other as was determined by transforming the *r*'s to *z* scores and testing for significance of difference. The WISC equalled the Binet in dependability for predicting itself over the four-year interval.

The data were further analyzed to find the intercorrelations of the various scales and to see if these differed at the two grade levels.

Table 3 shows that the abilities or characteristics measured by the Binet and WISC are ordered in essentially the same way for both grade levels. The largest discrepancy between correlations was found in the case of the WISC Performance and the Binet with the higher correlation occurring in the ninth

Table 3

Intercorrelations of IQ's and *t* Tests for Difference Between Correlations (Fifth Grade Followed by Ninth)

Test	WISC Performance	*t*	WISC Full Scale	*t*	Binet Form L	*t*
WISC Verbal	.44 .50	.43	.83 .86	.53	.78 .76	.27
WISC Performance			.86 .84	.37	.46 .64	1.39
WISC Full Scale					.73 .77	.48

grade. However, none of the differences between correlations was significant at the 5% level of confidence.

With regard to administration, Clarke (1) reported that the mean time for giving the Binet to the fifth grade was 76.8 minutes and for the WISC Full Scale 64.05 minutes, or a difference of 12.7 minutes. For the ninth grade the mean administration time was 69.2 minutes for the Binet and 55.9 minutes for the WISC, the difference being 13.3 minutes. At both levels, then, the WISC Full Scale averaged about 13 minutes less testing time than did the Binet.

Summary

Results of Binet and WISC administrations were compared, these tests having been given to 60 boys and girls in fifth grade and four years later to the same children in ninth grade. Both tests yielded IQ's that were relatively constant and showed equal stability over this interval of time as determined by comparisons between mean scores, test-retest correlations, and intercorrelations of major parts. On the average it required 13 minutes less to give a full WISC than to give a Binet at both grade levels.

Received August 1, 1955.

References

1. Clarke, F. R. A comparative study of the WISC and Revised Stanford-Binet, Form L, in relation to scholastic achievement of a fifth grade population. Unpublished master's thesis, Pennsylvania State Univer., 1950.
2. Delattre, Lois, & Cole, D. Comparison of the WISC and Wechsler-Bellevue. *J. consult. Psychol.*, 1952, **16**, 228–230.
3. Matyas, R. P. A longitudinal study of the Revised Stanford-Binet and the WISC. Unpublished master's thesis, Pennsylvania State Univer., 1954.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

Intolerance of Ambiguity and Ethnocentrism

Ronald Taft
University of Western Australia

In connection with a study on the autokinetic effect, a group of first-year psychology students at the University of Western Australia was observed in both individual and group situations.[1] Since these subjects had already been given a test of ethnic prejudice, an opportunity presented itself to test in another culture the previous findings in the U.S.A. on the relationship between prejudice and intolerance of ambiguity.

Following the work on this relationship by Frenkel-Brunswik (6), Block and Block (3) demonstrated that persons high on ethnocentrism were quick to develop personal norms in the extent of their perceived autokinetic movement. That is, in this unstructured situation, high ethnocentrics showed little tolerance for ambiguity. Similar findings using other material have been obtained by Fisher (5), O'Connor (7), and Siegel (8). Furthermore, Barron (1, 2) found that high ethnocentrics prefer simple rather than complex geometrical figures, and since persons who yield to social pressure also show this preference, we hypothesized that in a group autokinetic situation high ethnocentrics would tend to conform to their partner's judgments.

Method

The subjects consisted of 36 volunteers (21 females and 15 males). The measure of ethnocentrism was a Bogardus-type scale referring to the following groups: American Negroes, Chinese, English, Irish, Italians, Jews, Poles. Those whose average degree of ethnic aversion was greater than "Would admit to my social group as personal friends" were regarded as prejudiced, and all of the others as unprejudiced. There were 11 and 25 *S*s respectively in each group. Among the subjects used in this study prejudice is mostly not extreme and only five of the "prejudiced" subjects would have excluded more than two of the ethnic groups from citizenship in Australia. On the other hand, there were 13 extremely unprejudiced subjects who stated that they would be willing to accept members of nearly all groups even as marriage partners.

Autokinetic phenomenon. Each *S* was tested individually on the darkroom autokinetic effect. In the initial session they made 20 estimates of movement in the presence of *E*, a female graduate assistant. Unlike the customary procedure, the instructions included a short explanation of the autokinetic illusion; the *S*s were told that some people tend to see "a great deal of movement" and some tend to see "only a little" but almost everyone does tend to see movement of some sort. The fact that the *S*s were told that this movement was an illusion distinguishes this study from that described by Block and Block (3).

Approximately four weeks later, each *S* was tested similarly in the presence of some other *S*, estimates being made orally. Pairings were made of like sexes arranged so that persons seeing considerable movement in the preliminary sessions were paired with those seeing little movement.

Scoring was carried out as follows: (*a*) Intolerance of ambiguity. Whereas Block and Block used a rating method of determining when norms were established, we used an objective measure, the mean deviation. Because of the large range in the individual estimates, the relative variation was not a suitable measure. Since the instructions had informed the *S*s that there was actually no movement, the mean level of the estimates was used as a second measure of intolerance of ambiguity—the closer the estimates were to zero, the greater the intolerance. These indices were calculated for only the first ten judgments in the individual situation, in order to focus on *S*'s first reactions to the unstructured situation. (*b*) Conformity to partner. This was determined simply by the direction of *S*'s mean judgment in the group session in relation to his previous mean and that of his partner.

---

[1] The author thanks Mrs. Enid Cook for making available the data on the autokinetic effect.

Results

Intolerance of ambiguity. The more highly prejudiced *S*s made estimates that were closer to zero than did the unprejudiced. The mean estimates were 2.3 and 4.9 inches, respectively; using χ^2 with Yates's correction the difference was not statistically significant (p was actually .10). The mean *MD*'s of each individual's set of estimates were .8 and 1.0, respectively, also nonsignificant with a p of .10. However, when we combine these two indices we obtain results that are significant at the .05 level. Thus only one of the 11 prejudiced *S*s, compared with 12 out of the 25 unprejudiced, recorded both mean estimates and mean deviations above the median for the group. We seem justified therefore in concluding that the prejudiced *S*s are less able to tolerate ambiguity than the unprejudiced. A more detailed analysis of the results provides no indication that the relationship is other than linear between the very unprejudiced and the very prejudiced *S*s.

Conformity to social pressure. Results on the conformity index were available for only 21 *S*s (7 prejudiced and 14 unprejudiced), but the trend was a definite one. The unprejudiced *S*s tended to move *away from* rather than *toward* their partners. Only 6 out of 14 unprejudiced *S*s converged compared with 6 out of 7 of the prejudiced *S*s. This finding lends further support to the expectation that prejudiced persons are more dependent on their group, especially in an unstructured situation. Once more this trend was a linear one.

Conclusions

Further evidence is provided that those subjects who are comparatively higher on ethnocentrism are also more inclined to adopt an anchoring point quickly in an unstructured situation. The anchoring points used in the above experiment were zero movement, a steady personal norm, and the norms of one's partner. These results support Dana's generalized dictum "Ethnic prejudice is congruent with minimal resistance to the environment in which stereotypes are accepted, thinking is eschewed for conformity and security of absolutes" (4, p. 14).

Received June 27, 1955.

References

1. Barron, F. X. Some personality correlates of independence of judgment. *J. Pers.*, 1953, **21**, 287–297.
2. Barron, F. X. Complexity-simplicity as a personality dimension. *J. abnorm. soc. Psychol.*, 1953, **48**, 163–172.
3. Block, J., & Block, Jeanne. An investigation of the relationship between intolerance of ambiguity and ethnocentrism. *J. Pers.*, 1951, **19**, 303–311.
4. Dana, R. H. Personality orientation: an organizational focus for current research. *J. Psychol.*, 1954, **37**, 139–150.
5. Fisher, J. The memory process and certain psychosocial attitudes, with special reference to the law of Prägnanz. *J. Pers.*, 1951, **19**, 406–420.
6. Frenkel-Brunswik, Else. Intolerance of ambiguity as an emotional and perceptual personality variable. *J. Pers.*, 1949, **18**, 108–143.
7. O'Connor, P. Ethnocentrism, "intolerance of ambiguity," and abstract reasoning ability. *J. abnorm. soc. Psychol.*, 1952, **47**, 526–530.
8. Siegel, S. Certain determinants and correlates of authoritarianism. *Genet. Psychol. Monogr.*, 1954, **49**, 187–229.

Journal of Consulting Psychology
Vol. 20, No. 2, 1956

A WISC Profile for Retarded Readers

Grace T. Altus
Santa Barbara County Schools

This investigation posed the question: Is there a distinctive test pattern associated with the scores on the Wechsler Intelligence Scale for Children (WISC) for children with severe reading disabilities?

A severe reading disability was here defined as a discrepancy of two years or more between a given child's expected reading level as derived from his Full Scale WISC IQ and his actual reading level as measured by a standardized reading test. All children in the sample had been referred to the Guidance Department of the Santa Barbara County Schools by their teachers because of severe academic disabilities. All had Full WISC IQ's of 80 or more, spoke only English at home, had taken at least four subtests on each WISC scale, and were between third and eighth grade when given the reading test.

A group of 25 children from 12 elementary schools met these criteria. Twenty-four of them were boys—an exaggerated representation of the usual finding that boys outnumber girls as reading problems. The intelligence of the group was normal. Mean WISC IQ's were 97.8, 100.4 and 98.6 on the Verbal, Performance, and Full Scales respectively. There was less variability in the sample than in an unselected population as shown by standard deviations of 9.9, 10.3 and 9.2 IQ points on the same scales.

Since the mean Verbal-Performance IQ discrepancy was negligible, it was clearly not differentially diagnostic. However, the subtest patterning appears to be fairly distinctive, as shown in Figure 1. Coding and Arithmetic are significantly lower than Vocabulary, Digit Span, Picture Completion, Object Assembly, and Picture Arrangement at the .01 level of confidence. The Information subtest is significantly lower than Picture Completion at the .01 level and lower than Vocabulary and Digit Span at the .02 level of confidence. Had the positive correlations among the various subtests been taken into account in computing the significance of differences between subtest means, the chances of true differences would have been increased and Similarities would probably also have been included in the "low" subtests.

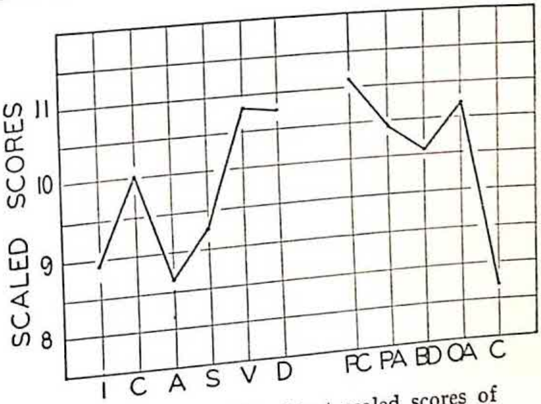


Fig. 1. Mean WISC subtest scaled scores of retarded readers.

The obtained WISC pattern is strikingly similar to the findings reported by Altus (1) regarding the differential validity of Wechsler-Bellevue subtests in predicting graduation of trainees from a camp for illiterate soldiers. Arithmetic, Information, and Digit Symbol (Coding) subtests were there shown to be highly effective in predicting graduation, i.e., achievement of technical literacy.

Thus the same subtests which are here significantly lower for children with severe reading disabilities were also lower for adult soldiers who failed to read at a specified level after instruction in reading. The consistency of the findings suggests that the profile herein described, while based on a small number of cases, may be reasonably characteristic.

Received August 17, 1955.

References

1. Altus, W. D. The differential validity and difficulty of subtests of the Wechsler Mental Ability Scale. *Psychol. Bull.*, 1945, 42, 238–249.
2. Wechsler, D. *The measurement of adult intelligence*. (3rd Ed.) Baltimore: Williams & Wilkins, 1944.

PSYCHOLOGICAL TEST REVIEWS

Test Reviews

Wechsler, David. *Wechsler Adult Intelligence Scale (WAIS)*. New York: Psychological Corporation, 1955.

Reviewed by

Roy Schafer

Yale University Medical School

The Wechsler-Bellevue Intelligence Scale (W-B) has had its face lifted. In form and substance, however, it is still very much with us in the Wechsler Adult Intelligence Scale (WAIS). The WAIS revision contains no new types of subtest. Its structure, like that of the W-B, makes possible not only pattern analysis of the equated subtest scores, but also comparison of Verbal and Performance IQ's, analysis of the sequence of passes and failures within each relatively homogeneous subtest as item difficulty increases, and qualitative analysis of the attempted task solutions.

As to whether the face-lifting has done much for the test, my impression is mixed; it is one of old wrinkles gone and new ones added and old charms gone and new ones added. Viewed generally, the WAIS remains our best clinical instrument for studying certain ego functions, such as judgment, concentration, and concept formation, as these operate in relatively well-structured, impersonal task situations. Clinical testers must study ego functions in such formalized and conventionalized situations in order to complete in each case their gross survey of personality organization and dynamics. Clinical experience with a battery of tests, including the Rorschach and TAT, repeatedly demonstrates that both versions of Wechsler's intelligence test contribute basic data both for locating significant areas of strength and weakness in defensive and adaptive functioning and for estimating the degree of strength or pathological impairment in these areas. Especially in studying borderline and acute psychotic conditions are these contributions important.

An outstanding improvement on the W-B is the all-new Vocabulary list containing mostly verbs, adjectives, and abstract or "literary" nouns (like *compassion*, *calamity*, and *fortitude*). The W-B Vocabulary contains more items (like *harakiri* and *guillotine*) overlapping Information. The WAIS Vocabulary more directly surveys the subject's means of verbal self-expression and through this it better defines his interest in verbal organization and communication of experience. On the other hand Information now contains two items that overlap Comprehension ("Why are dark clothes warmer than light-colored clothes?" and "How does yeast cause dough to rise?"); also Comprehension now contains three items that overlap tests of concept formation (three proverbs). Thus, not all the changes are in the direction of greater homogeneity within subtests.

Also commendable is the extension of the test's range in three important respects. (*a*) In most of the subtests, lower levels of intellectual function may now be tapped and scored through the inclusion of intellectually primitive lead-off items. Less often, the upper levels have also been raised. (*b*) The age range covered by the test has been extended upward by the establishment of norms through the ages 70–74 and even into the nebulous range of "75 years and over." Unlike the W-B, the WAIS does not cover the age range 10 through 15; the *Wechsler Intelligence Scale for Children* now does that job. In connection with matters of age, praise is due for the provision of tables whereby the weighted subtest scores may be, so to speak, corrected for age. This "correction" is achieved by comparing subtest performance with the average of the subject's age peers rather than comparing it with the average of the standard reference group, aged 20–34. Diagnostic scatter analysis in the older age ranges especially may well be furthered by these supplementary tables. (*c*) The third important extension is that the standardization population now includes Negroes and rural-agricultural persons and the sample was drawn from over the entire country. We may therefore apply the test with more confidence to our highly varied clinical populations. (It would be most interesting to see a breakdown of the data in terms of urban-rural, and Negro-White, as well as male-female.)

Most of the subtests have been lengthened. For clinicians this is a mixed blessing. On the one hand, greater length makes the test more time-consuming, and this increases the temptation to drop subtests or to use a short form of the test; on the other hand, greater length yields a broader sample of behavior in each subtest area and it increases score reliability. Increased score reliability, together with the WAIS's higher correlations between subtests and between single subtests and total test score, means on the average a narrowing of individual scatter of subtest scores. Consequently lesser amounts of scatter will now tend to be significant. However, this change may be more or less offset by the slight expansion of the weighted score scale (0–19).

The goal of eliminating ambiguous W-B items (p. 1) appears to account for the replacement of a number of items throughout the test. This revision too is a mixed blessing. A degree of ambiguity (or whatever makes for disruption of response—as in the frequent Information error that there are four pints in a quart) adds "projective" richness to intelligence test results. Ambiguity highlights the way repressive uncertainty limits and undermines intellectual attainment; it pinpoints confusional tendencies; it provokes obsessional rumination; it facilitates anxiety-laden temporary inefficiencies. To the clinician, misunderstandings and inefficiencies are not random events; they point to significant patterns of functioning. We must remember in this regard that there is no one standard for choosing items to study intelligence. In our lives our "intelligence" must cope with tasks involving all degrees of structure, all degrees of conventionality and familiarity, all degrees of complexity, all degrees of value-loading and conflict-relevance. It is in this sense that the Rorschach test becomes in part an intelligence test too, though it yields no precise score. Wechsler's psychometric approach to item selection almost necessarily favors as "well conceived" items that are easy and quick to administer and neatly scorable. But rounded clinical practice requires a description of more than IQ level; it concerns itself with intellectual potentialities, ambitiousness, "style," stress tolerance, and inhibitions. These intelligence characteristics can be culled only from the results of a battery of tests, and only when all test results are viewed from the standpoint of contemporary dynamic personality theory. In any event, the WAIS seems to be a more efficient machine than the W-B. Because of this, it seems reasonable to expect that temporary inefficiencies and other deviations of response will be even more diagnostically significant on the WAIS than the W-B.

Throughout the presentation of the WAIS it remains clear that Wechsler's approach to intelligence is limited by psychometric considerations. Dynamic considerations are underplayed. Concrete evidence for this is found in the crowded test blank which discourages verbatim recording and detailed qualitative notes. Similarly, the test manual completely ignores the many problems of technique so frequently encountered in clinical intelligence testing. What should we do when the subject too quickly gives up on items? How might we spot and follow up the implicit concreteness, misunderstandings, perceptual distortions, or bizarre conceptions often underlying responses? The WAIS manual even discourages clinical probing or testing the limits (p. 28). From the standpoint of getting an IQ that conforms neatly to the WAIS standardization, this discouraging attitude toward flexible and imaginative test administration is understandable. When, however, we use the test to help understand an individual, as well as to help estimate his general intelligence level, we need the standard administration as a firm guide rather than as a set of commandments; inevitably we must go beyond this guide, not rashly or frequently but with that sophisticated freedom that helps us study *response processes* and helps us clarify diagnostic and dynamic implications of logical or emotional disruption of responses. For example, anyone who has routinely asked subjects for the stories they have in mind on the last two Picture Arrangement items will know how often correct sequences are not backed up by a correct understanding of the stories, and vice versa.

While recognizing the practical difficulties in its way, I am sorry to see that no new type of subtest has been introduced. There is, for example, a crying need for a test of immediate memory for meaningful material; Digit Span and Digit Symbol are not satisfactory in this regard. It remains desirable, therefore, to supplement the administration of the WAIS with a Story Recall test such as that in Babcock's Test for measuring mental deterioration. Also unchanged—and unchallenged—is the view that the breakdown of the Full Scale IQ into "Verbal" and "Performance" components is adequate and satisfactory. Inspection of the tables of subtest correlations (pp. 15–17) makes it evident that Digit Span particularly and Arithmetic to a lesser extent do not cluster with the other four Verbal subtests. Also, performance on Digit Span and Arithmetic is relatively vulnerable to psychopathology. In deriving the Verbal IQ in clinical practice we therefore often get a better estimate of the patient's general verbal level by omitting pathologically lowered Digit Span and/or Arithmetic scores and then extrapolating. Another desirable change has not been made: the absence of part credits in the Block Design items continues to limit the sensitivity of scoring of that subtest. I wonder, by the way, why the colors on the blocks were changed. Such a change disrupts the continuity of our experience with test materials. I hope that the promised revision of *The Measurement of Adult Intelligence* will account for this change and for others that are not self-evident.

Clinicians will find it easy to switch from the W-B to the WAIS. Although the content of a fair number of items has been changed, the *type* of item remains the same throughout. Certain changes should be stressed, however: (*a*) some of the subtest instructions have been reformulated; (*b*) some of the criteria for scoring Information, Comprehension, and Similarities items have been altered; (*c*) the time limits and bonuses on performance items have been revised; (*d*) the IQ limits of the Defective and Very Superior ranges have been changed.

I must say in conclusion that a definitive clinical assessment of the WAIS must await two developments: (*a*) the appearance of Wechsler's revision of his book, for at the present time Wechsler knows so much more than the rest of us about the phenomenology of the revised test; (*b*) passage of enough time to allow the WAIS to be applied to a wide variety of subjects in routine clinical practice and in research projects. This review is thus no more than a record of my initial orientation while switching from the W-B to the WAIS. A follow-up appraisal of the WAIS five years from now seems definitely in order.

Wechsler, David. *Wechsler Adult Intelligence Scale (WAIS)*. New York: Psychological Corporation, 1955.

Reviewed by

Quinn McNemar

Stanford University

This revision (called WAIS) of the 1939 *Wechsler-Bellevue Intelligence Scale* retains the distinctive features of the original. About two-thirds of the earlier items have survived the rigors of item analysis and the over-all number of items has been increased from 217 to 257, but this increase is spread so unevenly over the subtests (23 of the 40 added items are Digit Symbol items) that the effective increase in total scale length is less than 10 per cent. The Vocabulary test, consisting of entirely new items, is an integral part of the Verbal and Full Scales.

The standardization sample of 1,700, ranging in age from 16 to 64, is claimed to be truly representative. It is a stratified sample with controls on sex, geographic region, urban-rural residence, race (white vs. nonwhite), occupational level, and education. A flaw in the sampling procedure was the (admitted) inability to select randomly within the strata. The reviewer shares the author's belief that any bias introduced by failure to follow the ideal random method is small. In other words, the reviewer regards the standardization sample as excellent, a marked improvement over that for the 1939 scale and over the norm groups for other scales of the same or earlier vintage.

A supplementary group of 352 cases ranging in age from 60 to beyond 75 provides the basis for norms at older age levels. Two claims are made for this group: it represents a good sampling from a typical American city (Kansas City, Mo.) and the test results for the 101 cases in the 60–64 age bracket are very similar to those of the same age in the main sample.

This time Wechsler has paid more attention to the problem of test reliability: split-half reliabilities are presented for each of the 11 subtests (except Digit Symbol) for each of three age groups. These reliabilities range from the .60's to the .90's, with a median value of .82. The reliabilities for Verbal, Performance, and Full Scale IQ's are reported as .96, .93, and .97, respectively. The reviewer is somewhat skeptical of these three coefficients, and hazards the prediction that test-retest reliabilities will not be as high.

Since WAIS is built on exactly the same base as its predecessor, it must be presumed that the principles underlying the old scale are being propagated. Stated differently, the author of WAIS shows no cognizance of the vast factor-analysis literature of the past 20 years. Regardless of whether one identifies with some sort of g school (typically British, and apparently adhered to by Wechsler) or with the primary ability (differential aptitude) approach, one cannot escape a lesson regarding test construction which emerges forcefully from the factor literature: the test constructor should strive for a pure measure of whatever he hopes to quantify. That is, a score should represent a point on a unidimensional scale rather than a hodgepodge of different dimensions.

Consider so-called Verbal IQ's obtainable on WAIS. Two persons with the same total weighted score on five of the tests can readily differ by as much as 10 points on the Digit Span or 6th test, thereby differing by 10 points on the Verbal IQ scale. Apparently, the user of WAIS is expected to condone such possible qualitative differences among Verbal IQ's, differences which are further complicated by the continued inclusion of the Arithmetic test as one of the measures of verbal ability.

On page 13 of the Manual a word of caution is injected regarding the possible unreliability of differences between scores on two subtests (same individual), but when we reach p. 18 we learn that "differences as large as five points may be unusual enough to be noteworthy." This statement, which is based on the statistical fact that the *median* value of the standard deviations of the 55 distributions of difference scores "was approximately three," ignores two other statistical facts; namely, the standard deviations (none of which is reported) vary from about 1.9 to about 3.6 and the reliability of the difference scores will vary considerably among the 55 sets of difference scores. And some of the reliabilities for difference scores will be disturbingly low (in the .50's and .60's). Since difference scores have been and are advocated for use in clinical diagnosis, the failure of the author to include in the Manual reliability coefficients and errors of measurement for difference scores is absolutely inexcusable.

In the opinion of the reviewer, the author of WAIS has attempted an impossible task: the construction of a scale to measure general (global) intelligence which at the same time will provide differences among subtests which are of diagnostic value. The eleven subtests are too diverse in content and yield intercorrelations too low to permit a satisfactory realization of the first aim, whereas they are not diverse enough (in factor-analytic sense) to yield the low intercorrelations necessary for reliable difference scores.

New Tests

Buhler, Charlotte, & Mandeville, Kathryn. *The Five Task Test (FTT)*. Ages 9–13. 1 form. (20) min. Record booklet ($7.50 per 50); set of colored pads ($4.00); Ball-and-field pad ($2.00); manual, pp. 24, mimeographed ($3.00); complete set ($15.00). Beverly Hills, Calif.: Western Psychological Services, 1955.

Buhler's "five tasks" are related performance tests, which are reported to have some relationships to emotionality and organic brain damage. The first three tasks are to cut a circle, a heart, and a star from four-inch squares of paper. The qualities of the products are rated against scaled specimens, and the number of scraps is counted as a measure of "emotionality." The fourth task is cutting "anything you wish" and is given a projective interpretation. The fifth task, the Terman ball-and-field test, is rated as normal, borderline, or problematic with the aid of scoring specimens. Like many other behavior samples, performance on these tasks may give a sensitive and alert psychologist many clues about a child's typical responses to his world. The evidence supporting these particular tasks as clinical instruments is weak. No data on reliability are given; the very moderate reported validities were tested against inadequately defined criteria. Perhaps the test is worthy of further research; perhaps it will remain a subjective instrument of some value to clinicians who depend on observations and hunches rather than on scores.—*L. F. S.*

Buhler, Charlotte, & Howard, Gertrude. *Personality Evaluation Form (PEF)*. 1 form. Booklet ($3.50 per 25) with manual, pp. 11, mimeographed. Beverly Hills, Calif.: Western Psychological Services, 1955.

Described by its authors as "a technique for the organization and interpretation of personality data," the *Personality Evaluation Form* is a 12-page booklet consisting mainly of a series of headings with blank spaces for entering data, inferences, and conclusions. As an outline for case study, the form has some merit. Like all such blanks, however, it lacks flexibility. In attempting to meet the requirements of all situations and all clients, it inevitably provides too little space for some topics and too much for other less relevant issues. Somewhat dangerous is the implication that it may be used by teachers with little psychological training. Well-trained clinicians hardly need such a blank; others should not attempt its use.—*L. F. S.*

Cooperative School and College Ability Tests (SCAT). College entrants and freshmen, 4 forms; grades 10–12, 2 forms. 70 (100) min. Test booklet ($3.25 per 25) with directions, pp. 14; hand-scoring answer sheets ($1.25 per 25), keys (35¢ set); IBM answer sheets ($1.25 per 25), keys (35¢ set); manual, pp. 57 ($1.00). Princeton, N. J., & Los Angeles, Calif.: Educational Testing Service, 1955.

The new Cooperative tests are an impressive contribution to the development of a group-administered battery designed for the limited but important function of educational prediction. Each test contains four parts—sentence completion, vocabulary, numerical computation, and numerical problem solving—selected for optimal concurrent validity from among nine types of items recommended by a panel of test specialists. Three scores are obtained, verbal, quantitative, and total, each expressed by a new type of three-digit scaled score and interpreted in terms of percentile norms. Currently available are two forms of Test 2 for use in the senior high school or with superior ninth-grade students, and four forms of Test 1, two of which have restricted distribution, for college entrants, freshmen, and sophomores. Further research is under way to produce additional levels, all with similar types of items, for use as far down the educational structure as the fourth or fifth grade.

Although the manual is modestly labelled as preliminary, it contains impressive data about the development of the tests. It is equally emphatic about their limitations. Kuder-Richardson reliabilities are .93, .91, and .95 for the verbal, quantitative, and total scores. A short form for survey purposes, containing two subtests and requiring only 50 minutes, has a reliability of .91 and correlates .97 with the longer test. Validities are presented in terms of concurrent correlations with marks in school subjects; predictive validities are to be obtained from future studies. A valuable feature is the emphasis on the confidence intervals of scores and a convenient graphic device for showing that a score represents a probable range of ability, not an absolute point. The provisional norms are based on a wide sampling of about 9,000 students in high schools and about 2,500 in colleges.

Few psychometric devices have equaled the new SCAT in the sophistication of the test construction research, and in the scope and clarity with which the data are presented—*L. F. S.*

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Response Set as a Suppressor Variable in the OAIS and MMPI

Benno G. Fricke[1]
University of Michigan

This article is concerned with a test taker's set to mark a particular response category "yes," "true," or "agree." According to Cronbach (5, 6), who has reviewed several response sets that influence a test taker's behavior, the variance generated by the response set discussed here is regarded as undesirable since, according to him, it contributes only error variance and cannot be used to increase the usefulness of a test. In this article a different rationale, supported by empirical evidence, will show how a measure of this response set can be used to improve the validity of many tests.

Nature of the Problem

Imbalance in scored responses. An inspection of responses scored in most personality tests or scales reveals that the significant responses are predominantly in one direction—i.e., "yes," "true," etc. For example, 96 of the 100 items in the Cornell Index are keyed "yes," and 47 of 60 items in the hysteria (Hy) scale of the Minnesota Multiphasic Personality Inventory (MMPI) are keyed "false." Some other tests in which a large majority of scored responses fall into a particular response category are the California Psychological Inventory, the Humm-Wadsworth Temperament Analysis, the Minnesota Teacher Attitude Inventory, and the Strong Vocational Interest Blank.

Response set as a suppressor variable. The contention here is that when the number of scored responses is unequally divided between all possible response categories the effect of a response set may be substantial. If response set per se is not directly related to the criterion, it operates to introduce error in the criterion-predictor. This is usually the case. If, however, response set is related to another variable that is related to the criterion, then response set may be used as a suppressor variable (14, 17, 18, 19, 21, 22, 25). By suppressing or removing the influence of response set from the criterion-predictor an improvement in test validity can be effected.

The Opinion, Attitude, and Interest Survey

A review (7) of personality tests validated to predict academic achievement discloses that for the more valid tests the response predictive of high achievement is usually false, no, or disagree. An imbalance was also found in an item analysis of the Opinion, Attitude, and Interest Survey (OAIS) a test constructed by the writer (7, 8). Results from the OAIS will be used to illustrate the major thesis of this paper.

Criterion group responses. The criterion groups used for the OAIS item analysis were 340 achieving and 473 nonachieving freshman men in the College of Science, Literature, and the Arts (SLA), University of Minnesota. The groups were equated for tested academic ability (ACE), and all had ACE percentile ranks above 54 on university freshman norms. Of the 126 discriminating statements only 33, or 26 per cent, are scored for high achievement if answered true. Why a negative reaction to some statements should be indicative of high achievement is difficult to understand. The statements in the OAIS are not concerned with study habits and attitudes. Typical statements and the response predictive of superior achievement are:

---

[1] This article was prepared in connection with a "group tests" course taught by the writer during the Winter Quarter, 1955, at the University of Minnesota as part of his Ford Foundation Teaching Intern Fellowship.

> Too much fuss is made over famous people (F); Often I wish I had more freedom (F); Most people have a very good imagination (F); I often find myself imitating someone I consider superior (F); People should go to church more than they do (F); I enjoy sleeping during the day more than at night (F); I'd rather be thought of as unusual than nervous (T); and Marriage before the age of twenty is usually foolish (T).

Interpretation and effect of imbalance. The writer originally held the view that the imbalance of true and false items was simply a function of how the statements were written. But this seems unlikely since on some tests (e.g., MMPI) having scales for several personality dimensions, the scored direction is different for different scales. The nature of the criterion groups, not the items, is probably responsible for the imbalance. The present interest is not in why criterion groups respond the way they do but in the effect of a strong response set. The specific question is "Can a measure of response set be used to increase test validity?"

It is obvious from the scored direction of statements in the OAIS that a test taker not reading the statements but marking false as the answer to all statements would receive a good score (a standard score of about 90). His high criterion-predictor score probably does not describe his "academic" personality. A test user viewing his score would usually overpredict his academic achievement. Another student having a "legitimate" response set for saying true would tend to mark an abnormal number of answers predictive of low academic achievement.

A measure of response set. One index or measure of set for answering "true" could be obtained by counting the times a person marked "true" to statements in a test. This method was used by Humm and Wadsworth (15) to get a measure of suggestibility or cooperativeness. They counted the times a test taker answered "no" to questions in the Temperament Schedule. Answer sheets having a very large or a very small "no count" are not considered sufficiently valid for further analysis. According to their suggested cutoff points, 30 to 60 per cent of the answer sheets would be rejected as invalid (1, 16).

A much more sensitive index of response set would be a count of the "true" responses for those statements which 49 to 51 per cent of the test takers mark true. For such statements there is no general agreement on an answer; they have maximum "controversiality." A person without a response set would respond true as often as false on these items. A person with a high score could be thought to have a strong set for marking true. A low score would indicate the opposite. It should be emphasized that items for such a "set" scale are selected independently of any criterion external to the test. In practice it is difficult to get enough items to form a scale at the 49–51 per cent level of controversiality. Accordingly, it was decided to accept for the "true" response-set scale, Set *T*, those statements in the OAIS which 40 to 60 per cent of each criterion group marked true. The Set *T* scale consists of 69 statements scored in the true direction. Typical Set *T* statements are:

> My friends consider me to be a talkative person; I do many things just to avoid criticism; Some people like to be considered rude and crude; I generally avoid domineering people; Others regard me as having a lot of patience; I'd rather talk than listen in a conversation; I am a very practical person; and Routine things require too much time.

Use of the Set T *scale as a predictor.* To determine whether Set *T* could be used as a suppressor variable, scores from it were correlated with the scores obtained from two grade-predictor scales constructed from an item analysis of the OAIS, and with the honor-point ratio (HPR) of freshmen in SLA. The first of the two grade-predictor scales, *S*, was based upon the familiar *single* response (true or false) item analysis; the second scale, *C*, was based upon a *configural* response item analysis which involves scoring simultaneously three responses (two content and one intensity) (7, 9). The statistics reported below were obtained from a cross-validation sample of 209 males who took the OAIS as part of the regular entering freshman test battery at least six weeks *before* the beginning of fall quarter classes.

Set *T* correlated − .036 with HPR, − .569

with S, and $-.505$ with C. The r of S with HPR was .326[2] and the r of C with HPR was .376.[3]

To see whether a combination of Set T scores and grade-predictor scores S and C would result in an increase in predictive efficiency, a multiple correlation coefficient (R) was computed.

The R of Set T and S with HPR was .394; the R of Set T and C with HPR was .427. The increase in both cases is appreciable.

Optimum weighting of suppressor and criterion predictor. It should be pointed out that the r of the corrected predictor with HPR will equal these R's only when the optimum amount of the suppressor, Set T, is added to the uncorrected predictors, S and C. If the suppressor and the uncorrected predictor are simply added, they are weighted according to their standard deviation. Several "trial and error" attempts will usually lead to an acceptable fraction of the suppressor to be added to the uncorrected predictor. Optimum weighting of suppressor and uncorrected scores probably should result in *no correlation between the suppressor and corrected scores.*

At first Set T scores were added to S and C, resulting in APS′ (academic personality-single response analysis), and APC′ (academic personality-configural response analysis). The SD's of Set T, S, and C were 7.8, 10.0, and 12.3, respectively. It will be noted that Set T is weighted more heavily in APS′ by this simple addition of scores. The r of APS′ with HPR was .351, and the r of APC′ with HPR was .404. Since these validity coefficients (r's) are lower than the R's obtained above it can be concluded that the grade predictors, S and C, were not weighted optimally with Set T. The r of Set T with APS′ was .25 and the r of Set T with APC′ was .15 which indicates that Set T was weighted too heavily. However when .7 Set T was added to S to form APS, the r of APS with HPR was .38; the r of .7 Set T with APS was $-.01$. When .8 Set T was added to C to form APC, the r of APC with HPR was .41; the r of .8 Set T with APC was .01.[4] For practical purposes, the influence of response set has been removed from the criterion-predictors as evidenced by the near-zero r with the suppressor. The important point is that a measure of response set, Set T, can be used to increase the accuracy of grade predictions, even though scores from it are not directly related to grade achievement.

Similarity of Set T of the OAIS and K of the MMPI

Construction of the K *scale.* Since the Set T scale of the OAIS functions as a suppressor variable, its relation to the K scale (18, 22) of the MMPI is of interest. It will be recalled that 22 of the 30 items in K were obtained by comparing responses of 50 psychopathic patients having a high L score and a normal MMPI profile, with responses of 604 normals (339 Minnesota men and women, and 265 college men and women) which were used in constructing other MMPI scales. The other eight items were selected from a study of faked responses (18). The 22 items will be referred to as L_6 items, as Meehl and Hathaway designated them, and the remaining eight items as "fake" items.

Structure of the K *scale.* An examination of the direction of scored responses in K reveals that 29 are false and only one is true,[5] the true item is from the L_6 items. A study of the percentage of true-false responses of the 604 normal cases[6] was made by averaging the percentage values for the 339 Minnesota and 265 college cases. Since not all normals responded true or false to each item, half the "?" responses were added to the true count. This crude analysis shows that 20 of the 30 items are answered true by 40 to 60 per cent of the normals. Without addition of half the "?" responses there are still 10 of the 30 items beyond the arbitrary limits. Only *one* of the L_6 items is not within a 36–64 per cent range. This item is: "I have very few quarrels with members of my family." Only 18 per cent of the normals give the "plus" response; this *K* item is scored *true*. Six of the eight "fake" items fall outside the 40 to 60 limits, but two (60.3 and 60.5 per cent) almost qualify.

---

[2] For the subgroup ($N = 138$) with ACE percentile rank above 54 the r was .41, and for the subgroup ($N = 71$) with ACE percentile rank below 55 the r was .10.

[3] The r's here were .44 and .20 for the high and low ACE subgroups respectively.

[4] The r of APS with ACE was .12; the r of APC with ACE was also .12.

[5] In this connection it is of some interest to find that of 40 items in Gough's personality scale to measure ability to make a good impression (11), 31 are scored false.

[6] The writer is grateful to Dr. S. R. Hathaway for supplying him the percentage frequencies for the two subgroups.

For practical purposes 22, or 73.3 per cent, of the items are in the controversial range; but strictly speaking 20, or 67.7 per cent, meet the requirement.

These percentages by themselves are not important if approximately two-thirds of the items in the MMPI are within the 40 to 60 per cent limits. An examination of 495 items on which data were available shows only 88, or 17.8 per cent, are found within the limits. Since 20 of the 88 are in the *K* scale, only 68 or 14.6 per cent of 465 non-*K* items are between 40 and 60 per cent. If half the "?" responses are not added, only 55, or 11.8 per cent, of 465 non-*K* items are within the limits. It is clear from this that the procedure used in constructing *K* has resulted in selecting items useful for measuring response set as described above in the OAIS analysis. It is possible that if a different psychopathic criterion group or a larger number of psychopathic patients were used for item analysis *more* than two-thirds of the *K* items would have been found in the controversial range.

Functional similarity of K *and Set* T. Further evidence on the similarity of Set *T* and *K* is found in their correlation with each other. The *r* of Set *T* with *K* for 209 freshman men is − .63; the *r* for 138 high ACE freshmen is − .58, and the *r* for 71 low ACE freshmen is − .71. The latter subgroup resembles the MMPI normal population more closely than the total SLA freshman group. The test-retest reliability of *K* for University of Minnesota freshmen was reported to be .51 by Berdie (2). Rosen (23) reported a .65 retest reliability for psychiatric patients. Hathaway and Monachesi (13) reported a retest reliability of .66 and Meehl and Hathaway (22) reported reliabilities of .72 and .74. It would appear that *Set* T *of the OAIS predicts* K *of the MMPI as well as* K *predicts itself.*

The *r*'s of Set *T* and *K* with several non-MMPI variables are presented in Table 1. The agreement in the *r*'s for uncorrected scores is quite strong. Minus *K*, rather than *K*, has been used to simplify comparisons. The Time score in the table is the time in minutes a test taker requires to complete the OAIS. The *I* (Infrequency) scale is similar to the *F* scale of the MMPI in that it is obtained by scoring those response configurations which are given by less than 4 per cent of each of the original criterion groups. HSRP is the test taker's high school rank percentile transformed to prohibit values. The other symbols are not new. The *r*'s of Set *T* and *K* with *S*, *C*, APS′, APC′, APS, and APC reveal a marked similarity in the function of Set *T* and *K*. The lower *r*'s of Set *T* and *K* with scales employing the configural items probably indicate that scores obtained from such scales are less subject to distortions because of response set, "defensiveness," etc.

Correlation of K *with clinical scales.* Table 2 presents some *r*'s of *K* with uncorrected MMPI scales. The consistently high negative *r*'s for *K* with *Pt*, *Sc*, and *Ma* are noteworthy. The *r*'s of *K* with *Pd* and *Hs*, though negative, are generally lower. The *r*'s with *Hy* are consistently positive and somewhat higher than the *r*'s for *Pd*. A generally negative but

Table 1

Correlations of Set *T* and *K* with Certain Non-MMPI Variables for 209 Freshman Men

	Variable												
Suppressor	HPR	Time	*I*	*S*	*C*	APS′	APC′	APS	APC	HSRP	ACE	Coop. Eng.	Ohio
Set *T*	−04	−17	−12	−57	−51	25	15	−01	01	00	−04	−06	−10
−*K*	−03	−09	02	−50	−38	00	03	−17	06	01	−03	−03	−05

Table 2

Correlations of *K* with Uncorrected MMPI Clinical Scales

			MMPI Scale								
Sample	Sex	*N*	*Hs*	*D*	*Hy*	*Pd*	*Mf*	*Pa*	*Pt*	*Sc*	*Ma*
Graduates (18)	M	100	−25	08	53	−09	−08	22	−65	−46	−42
College (24)	M	112	−31	−07	50	−09	−16	19	−59	−51	−30
College (4)	M	179	−48	−20	21	−26	−23	−12	−62	−57	−37
Normal (22)	M	100	−30	15	48	−17	—	−07	−67	−59	−36
Normal (22)	F	100	−35	−03	30	−06	—	−02	−64	−58	−28
NP patients (24)	M	110	−20	−19	14	−38	−19	−26	−66	−55	−45
NP patients (26)	M	100	−10	−04	15	−16	—	−15	−30	−25	−10
Abnormals (22)	M	100	−42	−29	11	−26	−08	−19	−60	−60	−37
Abnormals (22)	F	100	−17	−16	17	−21	04	−13	−63	−58	−38

low *r* is found for *Pa*, *Mf*, and *D*. The first five clinical scales mentioned above are those which are corrected by *K*. In view of the *r*'s with *Hy*, it would appear that a correction for it would be useful; more will be said on this later.

Correlation of K *and Set* T *with MMPI scales and optimum weighting*. Table 3 shows the *r*'s of Set *T* and *K* with five corrected and seven uncorrected scales. These *r*'s are especially interesting. The *r*'s for Set *T* and *K* are remarkably alike for all variables (the signs are opposite). From the correlations with *K* it would appear that *K* and *Pt* and *Sc* have been weighted reasonably well, as shown by the low *r* of *K* with *K*-corrected *Pt* and *Sc*; perhaps *Hs* and *Pd* have been overweighted and *Ma* underweighted with *K*. *These are also the conclusions to be drawn from the* r's *of Set* T *with the* K*-corrected scales*. The *r*'s of *K* with uncorrected scales are similar to those in Table 2. The *r*'s of Set *T* with uncorrected scales are also similar to the *r*'s in Table 2. It would appear that Set *T* and *K* accomplish essentially the same thing. Whether Set *T* is superior to *K* is not determinable from the available data. While *K* may not be as pure a measure of response set, it may do something in addition to what is done by Set *T*.

Prior to collecting all the data involving *K* the writer suggested to Hathaway that perhaps .4*K* or .5*K* should be added to *Ma* rather than .2*K*. Hathaway located in his files the differential ratio (18) curve for the sample used to determine how much *K* should be

Table 3

Correlations of Set *T* and *K* with Five *K*-corrected and Seven Uncorrected MMPI Scales

				MMPI Scale											
Suppressor	Sample	Sex	*N*	*L*	*F*	*Hs* +.5*K*	*D*	*Hy*	*Pd* +.4*K*	*Mf*	*Pa*	*Pt* +1.0*K*	*Sc* +1.0*K*	*Ma* +.2*K*	*Si*
Set *T*	College freshmen (7)	M	209	−23	27	−26	02	−28	−17	20	04	02	02	32	25
K	College freshmen	M	209	44	−41	42	03	36	26	−23	−09	−05	12	−31	−49
K	Ninth graders (13)	M	200	52	−42	45	08	45	27	−05	−14	11	08	−23	−27
K	Ninth graders (13)	F	200	44	−44	22	07	32	11	03	−08	−01	−07	−25	−52

added to *Ma*. This curve showed the validity of *Ma* plus *K* increased rapidly to *.2K*, remained at this level until *.5K*, then decreased. Hathaway recalled that *.2K* was selected to avoid adding additional error variance. This writer believes in view of the data collected to date that the addition of *.4K* or *.5K* would improve the validity of *Ma*. As mentioned above the *r*'s of Set *T* and *K* with *Hs* and *Pd* in Table 3 are offered as evidence that too much *K* has been added to *Hs* and *Pd*. Additional evidence that the *Pd* scale may be overcorrected is found in a reanalysis of Capwell's data (3). McKinley, Hathaway, and Meehl (18) reported the *K*-corrected *Pd* scale decreased the differentiation of Capwell's adolescent delinquents from their matched controls.

The MMPI: Operation of the *K* Scale

The manner in which *K* operates to improve the validity of certain MMPI scales has not been described in the literature. One way of seeing how *K* works is through an analysis of the scored responses for each clinical scale. It should be obvious that focusing attention on the scored direction of responses is not a very precise method of analysis. This approach ignores the percentage of clinical and normal persons responding to each item; one item to which a large number of one or both groups respond in the scored direction may count as much as several items to which a very small number of one or both groups respond in the scored direction: e.g., half the *Hs* items are answered in the plus direction by less than 15 per cent of the normal population. Other factors such as unusual wording of items may contaminate the results: e.g., many *Hs* items are worded in the negative, "I do not often notice my ears ringing or buzzing." The effect of items which are common to several scales is another important factor. Wheeler, Little, and Lehner (24) point out, for example, that an *r* of + .46 can be expected between *Hs* and *Hy* because of 20 overlapping items. And perhaps most important, response set, as Cronbach (5, 6) points out, is probably of little consequence when the items are unambiguous.

Direction and significance of scored responses for uncorrected scales. It was mentioned earlier that the items in different MMPI scales are scored in different directions. Table 4 presents the number and percentage of scored responses as well as the standard score a test taker would have if all items were marked true, or if all items were marked false (12). The magnitude of the differences in standard scores and percentage of true and false should be noted. The "true response" profile increases from left to right and resembles what has become known as a psychotic-type profile, the "false response"

Table 4

The Number and Percentage of True-False Scored Responses and Standard Scores (for Men) for 13 Uncorrected MMPI Variables

Scored response	MMPI Scale *L*	*F*	*K*	*Hs*	*D*	*Hy*	*Pd*	*Mf*	*Pa*	*Pt*	*Sc*	*Ma*	*Si*
True	0	44	1	11	20	13	24	28	25	39	61	35	35
False	15	20	29	22	40	47	26	32	15	9	17	11	35
Total	15	64	30	33	60	60	50	60	40	48	78	46	70
Percentage true	0	69	3	33	33	22	48	47	63	81	78	76	50
Percentage false	100	31	97	67	67	78	52	53	37	19	22	24	50
Percentage difference	−100	38	−94	−34	−34	−56	−4	−6	26	62	56	52	0
T score when all marked true	50−	80+	29	65	58	44	75	65	100	91	120	97	62
T score when all marked false	80+	80+	81	90	106	106	81	73	70	49	60	43	62
T score difference			−52	−35	−48	−62	−6	−8	30	42	60	54	0

Table 5

The Number and Percentage of True-False Scored Responses and Standard Scores (for Men) for Five *K*-corrected MMPI Clinical Scales

	MMPI Scale				
Scored response	*Hs* +.5*K*	*Pd* +.4*K*	*Pt* +1.0*K*	*Sc* +1.0*K*	*Ma* +.2*K*
True	12	25	40	62	35
False	37	38	38	46	17
Total	49	63	78	108	52
Percentage true	24	40	51	57	65
Percentage false	76	60	49	43	35
Percentage difference	−52	−20	2	14	30
T score when all marked true	52	64	85	120	96
T score when all marked false	116	95	81	96	50
T score difference	−64	−31	4	24	46

profile decreases from left to right and resembles the neurotic-type profile.

A scale to measure response set is likely to be useful when the discrepancy in percentage of true-false is greater than, say, 40 (the discrepancy was 48 per cent on Scale *S* of the OAIS). Application of this discrepancy conceivably could result in improving 4 of 10 MMPI clinical scales. The three scales having the majority of scored responses as true (*Pt*, *Sc*, and *Ma*) could be adjusted by *adding*, say, Set False scores and the one scale having most scored responses as false, *Hy*, could be adjusted by *subtracting* Set False scores. The fraction of Set False to be added to, or subtracted from the clinical scale can be determined empirically as described earlier for Set *T* of the OAIS.

Of course since only the scored direction is being taken into account, a measure of response set will not necessarily improve the validity of these scales. A correlation, positive or negative, between Set False and a clinical scale probably indicates that Set False could be used as a suppressor. If Set False (or for that matter *any* variable) is not related to the clinical behavior, but is related to the MMPI clinical scale, then Set False may be used as a suppressor even though there is no true-false imbalance.

If *K* is accepted as the measure of response set in place of Set False, to add false responses to *Pt*, *Sc*, and *Ma*, and subtract false responses from *Hy*, the validity of the four clinical scales would be improved. It will be recalled (Tables 2 and 3) that the *r*'s for the uncorrected clinical scales with *K* were largest for *Pt*, *Sc*, *Ma*, and *Hy*.

Interpretation of K *scale*. Meehl and Hathaway (22) reported that *r*'s between *K* and a clinical scale "as low as .20 can be utilized to yield very significant and useful improvements in discrimination." The validity of *Pt*, *Sc*, and *Ma* was improved by the addition of *K* or a fraction of *K* (18). It appears (18, 22) that attempts were made to *add* *K* to *Hy*. That they should attempt addition of *K* and *Hy* follows from the commonly held view of the *K* scale. McKinley, Hathaway, and Meehl (18) wrote, "since high *K* scores represent the defensive or fake good end of the test attitude continuum, the most obvious approach to the problem is to add *K* (or some fraction of *K*) to the raw score on each personality variable, i.e., increase the score in the direction of abnormality." They apparently felt the *K* correction for *Hy* resulted in no improvement because the *Hy* scale contained 10 of the 30 *K* items which were scored in the same direction. If *K* is viewed as a measure of response set, then *K* or some fraction of it should be *subtracted* from *Hy*. That is, a person with a high *K* has probably earned a higher *Hy* than he should have. To say that he has "faked bad" is debatable. If *K* is not an adequate measure of response set, then false responses from a scale such as Set F could be subtracted from Hy, or true responses from a Set T scale could be added to Hy.

Direction and significance of scored responses for K*-corrected scales.* Table 5 shows data similar to those found in Table 4 except that the K-corrected scales are described. On the basis of the number of true and false scored responses in Table 5 the r's for Set T and K with the K-corrected scales in Table 3 are to be expected: e.g., the r between K and $Hs + .5K$ is positive, the r between K and $Ma + .2K$ is negative, and the r between K and $Pt + 1.0K$ is about zero.

For all K-corrected scales in Table 3 the r's with K are consistently less negative (some are positive) than the comparable r's in Table 2. From Tables 4 and 5 it can be seen that these changes in r are accompanied by a consistent increase in the proportion of false responses in the clinical scales. However the r's of Tables 2 and 3 indicate that the true-false imbalance shown in Tables 4 and 5 will not explain all the r's obtained for the suppressors K and Set T with the clinical scales: e.g., uncorrected Hs correlates negatively with K despite the fact that 67 per cent of the 33 Hs items are scored false; Si correlates positively with Set T and negatively with K despite the fact that 50 per cent of the Si items are scored false. Several possible explanations for these unexpected correlations have been mentioned. Perhaps the important point here is that a measure of response set such as K or Set T may function as a suppressor when there is no true-false imbalance, and may not function as a suppressor when there is a moderate true-false imbalance.

Correlations of K *with other MMPI scales.* Of considerable interest is the relation of K to several nonclinical scales which were constructed for the MMPI and reported by Meehl and Hathaway (22). K was found to correlate $-.70$ with N, a scale derived empirically by Meehl (20) to correct for self-criticality of certain normal plus-getters who show deviant profiles. As would be expected, the majority (50 of 78) of the items are scored true. The r of K with G and + (plus) was $-.76$ and $-.64$, respectively; both these scales were derived wholly by a method of internal consistency and without regard to nontest behavior. Only 9 of 62 items in G, and 11 of 55 items in + are scored false.[7] The r of K with C_H, the correction factor for the old hypochondriasis scale (H-C_H), was $-.67$; C_H was derived from an item analysis of the responses of the original 50 hypochondriacal criterion cases and 50 nonhypochondriacal psychiatric patients who scored high on H.[8] Of the 48 items in C_H only 13 are scored false. The r of K with Hy-O was .81; the Hy-O scale consists of the zero items in the Hy scale and 19 of its 20 items are scored false. Meehl and Hathaway reported a factor analysis of a correlation matrix of these special scales revealed "one common factor is quite sufficient to account for the intercorrelations of these scales." *It may be that this common factor is response set,* and that the best measure of it might be obtained in the manner described for Set T.

Some Implications for Personality Test Research

Studies need to be done on the *nontest behavior of those with a strong or weak response set* to mark a particular response alternative. The determination as to what is measured by scales such as Set T and K should be made in this way. Too frequently *scores* from the K scale have been correlated with *scores* from other tests, and judgments have been made as to what K was *really* measuring. Validity indicators such as L and F should be composed of items without a true-false imbalance so that independent measures are obtained.

In view of the ease with which an *OAIS* Set T-type scale can be constructed, the writer recommends that response set scales be prepared for other empirically validated personality tests; scales correlating with response set should be corrected. It seems unlikely that response set keys would be able to bring validity to tests which are not validated empirically even though response set may be a major component in a test taker's score. It is important that those who derive and use empirically validated personality tests be concerned not only with the degree to which an item discriminates between criterion groups but also with the *direction* of the scored responses. A recent review by Furst and Fricke (10) discloses that rarely does a scale constructor indicate the direction of the scored responses. The correlations between many scales and tests can be "explained," at least in part, by an inspection of the scored responses and recognition of the role that response set may have played.

---

[7] The writer obtained the scoring stencil for G and + from Dr. S. R. Hathaway to check the inference that most of the items were scored true.

[8] The construction of the present 33 item Hs scale has not been described in the literature; it consists of 31 items taken from the 55 item H scale, and two items which were not in the old hypochondriasis (H-C_H) scale.

Summary

A method for constructing a scale to measure a test taker's set to say "true" to personality test items was described. The Set T scale of the Opinion, Attitude, and Interest Survey (OAIS) was used to show how response set could be harnessed to function as a suppressor variable and improve the validity of two empirically validated grade-predictor scales.

The marked structural and functional similarity of Set T of the OAIS and K of the MMPI was drawn upon to challenge the traditional interpretation of the K scale. Some evidence was assembled which indicates that some of the MMPI scales are not optimally K-corrected.

It was suggested that those who construct and use personality tests should be concerned with the *direction* of the scored responses as well as with the degree to which an item discriminates.

Received August 29, 1955.

References

1. Arnold, D. A. The clinical validity of the Humm-Wadsworth Temperament Scale in psychiatric diagnosis. Unpublished doctor's dissertation, Univer. of Minnesota, 1942.
2. Berdie, R. F. *Test-retest correlations of the MMPI for freshmen at the University of Minnesota.* Student Counseling Bureau, Univer. of Minnesota, 1951.
3. Capwell, Dora F. Personality patterns of adolescent girls: II. Delinquents and non-delinquents. *J. appl. Psychol.*, 1945, **29**, 289–297.
4. Cottle, W. C., & Jackson, O. P. The effect of random answers to the MMPI. *Educ. psychol. Measmt*, 1951, **11**, 224–227.
5. Cronbach, L. J. Response sets and test validity. *Educ psychol. Measmt*, 1946, **6**, 475–494.
6. Cronbach, L. J. Further evidence on response sets and test designs. *Educ. psychol. Measmt*, 1950, **10**, 3–31.
7. Fricke, B. G. The development of an empirically validated personality test employing configural analysis for the prediction of academic achievement. Unpublished doctor's dissertation, Univer. of Minnesota, 1954.
8. Fricke, B. G. *The Opinion, Attitude and Interest Survey.* Minneapolis: Investors Diversified Services, 1955.
9. Fricke, B. G. A configural-content-intensity item for personality measurement. *Educ. psychol. Measmt*, 1956, **16**, 54–62.
10. Furst, E. J., & Fricke, B. G. Development and applications of structured tests of personality. *Rev. educ. Res.*, 1956, **26**, 26–55.
11. Gough, H. G. On making a good impression. *J. educ. Res.*, 1952, **46**, 34–35.
12. Hathaway, S. R., & McKinley, J. C. *Manual for the MMPI.* New York: Psychological Corp., 1951.
13. Hathaway, S. R., & Monachesi, E. D. *Analyzing and predicting juvenile delinquency with the MMPI.* Minneapolis: Univer. of Minnesota Press, 1953.
14. Horst, P. The prediction of personal adjustment. *Social Science Research Council Bull.*, 1941, No. 48.
15. Humm, D. G., & Wadsworth, G. W. *Manual for the Humm-Wadsworth Temperament Scale.* Los Angeles: Author, 1935.
16. Humm, D. G. Discussion of Dorcus' study of the Humm-Wadsworth Temperament Scale. *J. appl. Psychol.*, 1944, **28**, 527–529.
17. Levine, A. S. A technique for developing suppression tests. *Educ. psychol. Measmt*, 1952, **12**, 313–315.
18. McKinley, J. C., Hathaway, S. R., & Meehl, P. E. The Minnesota Multiphasic Personality Inventory: VI. The *K* scale. *J. consult. Psychol.*, 1948, **12**, 20–31.
19. McNemar, Q. The mode of operation of suppressant variables. *Amer. J. Psychol.*, 1945, **58**, 554–555.
20. Meehl, P. E. An investigation of a general normality or control factor in personality testing. *Psychol. Monogr.*, 1945, **59**, No. 4 (Whole No. 274).
21. Meehl, P. E. A simple algebraic development of Horst's suppressor variables. *Amer. J. Psychol.*, 1945, **58**, 550–554.
22. Meehl, P. E., & Hathaway, S. R. The *K* factor as a suppressor variable in the MMPI. *J. appl. Psychol.*, 1946, **30**, 525–564.
23. Rosen, A. Test-retest stability of MMPI scales for a psychiatric population. *J. consult. Psychol.*, 1953, **17**, 217–221.
24. Wheeler, W. M., Little, K. B., & Lehner, G. F. The internal structure of the MMPI. *J. consult. Psychol.*, 1951, **15**, 134–141.
25. Wherry, R. J. Test selection and suppressor variables. *Psychometrika*, 1946, **11**, 239–247.
26. Williams, H. L., & Lawrence, J. F. Comparison of the Rorschach and MMPI by means of factor analysis. *J. consult. Psychol.*, 1954, **18**, 193–197.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Two Measures of Anxiety[1]

Erwin J. Lotsof and Walter L. Downing
University of California, Los Angeles

The chronic anxiety syndrome, as defined by Cameron, is "characterized by the presence of persistently heightened skeletal and visceral tensions . . ." (2, p. 249). Using this definition, Taylor (5) devised a paper-and-pencil test to measure manifest anxiety. Mowrer (4) and Bixenstein (1) use the term *tension* in the therapeutic situation in a way which appears to be synonymous with anxiety. They measure changes in tension by recording palmar perspiration, and by computing a Discomfort-Relief Quotient. In view of the apparent similarity of these two concepts, anxiety and tension, it was felt that the two measures, if applied to the same population, should show a substantial correlation.

The Taylor Scale of Manifest Anxiety (TS) was administered to an introductory psychology class. A week later, 30 of the students participated in another "experiment" in which palmar perspiration (PP) was recorded by the method described in Mowrer (4). No indication was given the participants that the two measures were related, nor that they concerned anxiety. Two judges ranked the fingerprints for density in an almost identical sequence (rho = .99). The range of TS scores was 1 to 26, with a mean of 10.23, and an *SD* of 6.33. The TS scores were ranked, and the two measures gave a rank-order correlation of 0.00.

Although the number of *S*s was small, some correlation would be expected if any relationship existed between the variables being measured. If anxiety is conceived of as "permeating" the entire personality, and if tension is conceived of as a response of the total organism, then the results are difficult to explain.

Assuming that the tests are valid measures of manifest anxiety and of tension, two hypotheses might be offered to account for the findings. (*a*) Anxiety and tension are intervening variables whose consequences become manifest in certain situations. The negative findings are due to the different testing times. (*b*) The terms have different sets of referents. The test of anxiety is psychological; that for tension is physiological. Although there is a large semantic overlap, the referents may be concerned not with different levels of the same event, but with different events.

It would appear desirable, therefore, in using concepts of anxiety and tension, to follow Maslow (3) in indicating the source of the concept with a subscript.

Brief Report.
Received February 23, 1956.

References

1. Bixenstein, V. E. A case study of the use of palmar sweating as a measure of psychological tension. *J. abnorm. soc. Psychol.*, 1955, **50**, 138–141.
2. Cameron, N. *The psychology of behavior disorders.* Boston: Houghton Mifflin, 1947.
3. Maslow, A. H. A suggested improvement in semantic usage. *Psychol. Rev.*, 1945, **52**, 239–240.
4. Mowrer, O. H. (Ed.) *Psychotherapy: theory and research.* New York: Ronald, 1953.
5. Taylor, Janet A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.

---

[1] An extended report of this study may be obtained without charge from Erwin J. Lotsof, Dept. of Psychology, University of California, Los Angeles 24, Calif., or for a fee from the American Documentation Institute. Order Document No. 4827 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Social Desirability in the MMPI[1]

Wilbert E. Fordyce
Veterans Administration Hospital, Seattle, Washington

The concept of social desirability (hereafter, *S-D*) has been explored and elaborated in several studies (1, 2, 3, 4, 6). The concept concerns elucidating a basis for describing and controlling test-taking attitudes on personality inventories and schedules. A preliminary definition is, "Consensus judgments as to what behavior, feelings, and attitudes win social approval in American society." A more operational specification will be given below.

Problem

The impetus for studying *S-D* stems largely from a study by Edwards (1) in which he found frequency of endorsement of personality inventory items as self-descriptive correlated .87 with the rated social-desirability values of the items. Edwards proceeded to construct the Personal Preference Schedule (3), an objective-type personality test, based on Murray's need system, in such a way as to control this test-taking attitude of endorsing socially desirable items. The construction of the test involves the assumption, supported by the studies cited above, that a major part of the variance in these types of tests is accounted for by the test taker's desire to subscribe to characteristics that are desirable in our society. The present study concerns itself with an extension of this concept of *S-D* to the Minnesota Multiphasic Personality Inventory (MMPI) (5).

The principal validity scales of the MMPI (*F* and *K*) attempt to characterize the subject's approach to the test and his readiness to subscribe to items in the clinical scales descriptive of mental illness. The relationships between these validity scales and the clinical scales are formalized in the instance of *K* by applying a correction to five of the clinical scales based on the *K* score. With the general notion that the subject's approach to the test can be substantially characterized by his readiness to respond to socially desirable or socially undesirable items, and that this approach is reflected in both the validity and clinical scales of the MMPI, the hypotheses for this study are:

a. There will be a significant correlation between an *S-D* scale and the *F* and *K*[2] scales of the MMPI.

b. There will be a significant correlation between an *S-D* scale and the clinical scales of the MMPI.

Method

The measure of *S-D* used in this study defines the concept. A subset of MMPI items consisting of those making up the *F*, *K*, and Taylor Anxiety (7) scales was presented to a group of 10 judges. The judges ranged from faculty to students and departmental secretaries of the University of Washington Department of Psychology.[3] They were asked to select the items, and the direction of their scoring, which would be socially desirable responses when endorsed by test subjects. Only those items on which all 10 judges agreed as to selection and direction of scoring (true or false) made up the *S-D* scale. Seventy-nine items met this criterion. The *S-D* score became the number of these 79 items subjects answered in the appropriate direction.

Subjects consisted of all the available, completed, group-form MMPI records on male

---

[1] From Veterans Administration Hospital, Seattle, Washington, and Department of Psychiatry, University of Washington School of Medicine.

[2] Edwards (3) has already reported a correlation of .63 between *S-D* and *K*.

[3] The scale was constructed by Allen L. Edwards, University of Washington.

Table 1

Item Overlap Between *S-D* Scale and Other MMPI Scales

No. of items	MMPI Scale											
	L	*F*	*K*	*Hs*	*D*	*Hy*	*Pd*	*Mf*	*Pa*	*Pt*	*Sc*	*Ma*
No. in MMPI Scale	15	64	30	33	60	60	50	60	40	48	78	46
Scored same in *S-D*	0	0	5	0	2	1	2	0	0	0	0	3
Scored opposite in *S-D*	0	48	1	6	9	7	4	0	11	5	11	2

psychiatric and/or neurological patients in the VA Hospital, Seattle. Only the first MMPI record obtained in this hospital was used. Records with more than 25 items omitted were excluded. There were 97 records meeting these criteria. Since testing occurs on a consultation basis, this sample is not a random representation of the neuropsychiatric population in this hospital. However, except for the exclusion of patients too hyperactive, confused, or with motor disabilities prohibiting handling the test, the sample would appear to be reasonably representative of hospitalized male veteran psychiatric and/or neurological patients.

Results and Discussion

Since both hypotheses involve the degree of correspondence between the *S-D* scale and the regular MMPI scales, the question of influence of item overlap becomes important. This issue is particularly important with regard to the first hypothesis, which raised the question of the amount of correlation between *S-D* and *F* and *K*, because the *S-D* scale was constructed from an item pool based on these scales and the Taylor Anxiety scale. Table 1 shows the amount of item overlap between the *S-D* scale and each of the regular MMPI scales.

It seems evident from inspection of Table 1 that item overlap is not sufficient to establish substantial correlations between *S-D* and the validity and clinical MMPI scales, except for *F*. In that instance the 75 per cent overlap would assure a substantial correlation between the two only as long as the direction of scoring of the overlapping items was consistent. The fact that all 48 *F* items occurring in *S-D* are scored in a single direction on *S-D*, by unanimous judgment of 10 judges, is itself indicative of a correlation of *F* with *S-D*. In order to be able to evaluate separately the effects of *S-D* and *F*, the item overlap between the two scales was eliminated in a subscale of *S-D*, by scoring all records separately for the 31 *S-D* items not common to *F*. This subscale is designated as, *S-D-F*.

Table 2 gives the correlations between *S-D* and *S-D-F* and the clinical and validity scales of the MMPI, and between *F* and *K* and these scales for the 97 test records. Table 2 indicates *S-D* has a highly significant correlation with both *F*, − .82, and *K*, .69, confirming the first hypothesis. The fact that *S-D* correlates higher with each of these scales than they do with each other (− .55) seems to indicate at least one common factor underlies *F*, *K*, and *S-D*, and that *S-D* better estimates this common factor than does either

Table 2

Correlation of *S-D* and Validity Scales with Clinical and Validity Scales, Without *K* Correction

MMPI Scales	Comparison Scales			
	S-D	*S-D-F*	*F*	*K*
L	24*	26**	10	30**
F	−82**	−60**		−55**
K	69**	71**	−55**	
Hs	−70**	−49**	51**	−60**
D	−69**	−70**	56**	−45**
Hy	−38**	−47**	25*	−13
Pd	−57**	−33**	56**	−41**
Mf	−21	−33**	35**	−23*
Pa	−60**	−47**	62**	−32**
Pt	−86**	−65**	69**	−75**
Sc	−91**	−62**	84**	−70**
Ma	−50**	−30**	49**	−50**
Mean†	−651**	−498**	566**	−479**

* Significant departure from zero at 5% level.
** Significant departure from zero at 1% level.
† Nine clinical scales only.

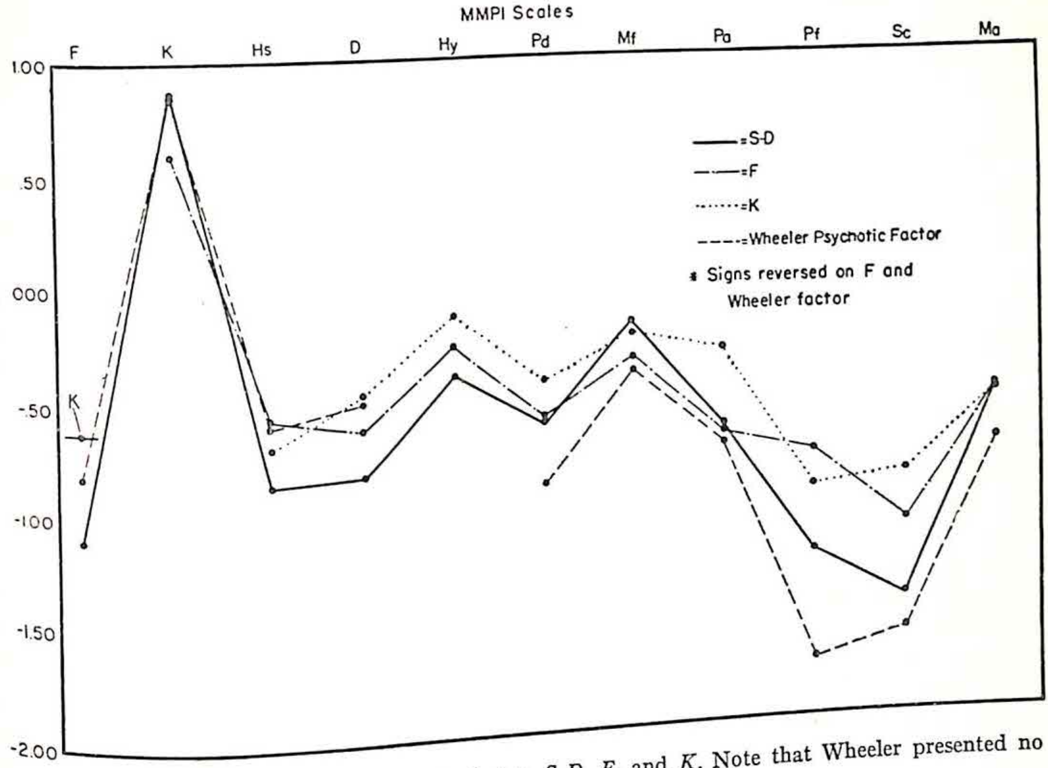


Fig. 1. *Z′* values of the Wheeler psychotic factor, *S-D*, *F*, and *K*. Note that Wheeler presented no loading of his factor on *Hy*.

F or *K*. The correlation of − .60 between *S-D-F* and *F* clearly indicates that when the *F* items are removed from *S-D* there is still a highly significant relationship.

The second hypothesis, that *S-D* will correlate significantly with the clinical scales, is also substantiated by the correlations presented in Table 2. The mean correlation between *S-D* and the nine clinical scales of − .651 is very highly significant. Only the *Mf* scale has a correlation which is not significant. The mean correlation of *S-D-F* with the clinical scales of − .498 is also highly significant, again indicating that *S-D* is something more than the *F* scale. Since *S-D-F* at least roughly approximates *S-D* in size and pattern of correlations in Table 2, in the interest of brevity it will receive no further consideration.

The pattern or profile of correlations in Table 2 is quite similar in shape for *S-D*, *F*, and *K*. This lends further support to the notion that there is an underlying factor common to those three scales. Since the correlations between these three scales and the clinical scales are so high, it seems evident the factor is well reflected in several of the clinical scales. At this point a study was made of one factor analysis of MMPI records to investigate further this apparent underlying factor.

Wheeler, *et al.* (8), using a sample of 112 male veteran hospitalized psychiatric patients, intercorrelated and factor analyzed a pool of MMPI records. This resulted in two factors labeled, *Psychotic* and *Neurotic*. For the purposes of this study only the first Wheeler factor will be considered. This Psychotic factor was characterized thus: ". . . when this factor is present to a marked degree in an individual, the usual ego-defensive mechanisms are held in abeyance and the person now tends to show himself in the worst possible light" (8, p. 139). This obviously resembles a definition of social *un*desirability. If that is the case, the loadings of the clinical scales on this factor should resemble the distribution of correlations between the *S-D* scale and the clinical scales. Since Wheeler's factor loadings were obtained by orthogonal rotations, it be-

Ednl. Psy. Research
Dated ....
Acc. No ....

comes feasible to treat his factor loadings as correlations, permitting direct comparison with the correlations presented in Table 2. In order to correct for the skewed distribution of *r* and pull apart the distributions of *F, K, S-D,* and the Psychotic factor on the higher correlations, these data are presented as *z'* transformations of the original correlations. They are presented in Figure 1 in graphic form, with the signs of *F* and the Psychotic factor changed to negative to make visual comparison easier.

The striking thing about Figure 1 is the remarkable similarity in shape of the four distributions, *F, K, S-D,* and the Psychotic factor. It seems evident that each of the four is well saturated with some common component. It will be noted from Figure 1 that *S-D* more closely approximates the shape and elevation combined of Wheeler's factor than does either *F* or *K*. This greater conformity of *S-D* is indicated by the sum of the squares of the differences of *z'* values between each of *S-D, F,* and *K* with the Psychotic factor across the nine clinical scales. This crude measure of conformity yielded values of 1.53 for *K*, 1.03 for *F*, and .52 for *S-D*. The smaller value for *S-D* clearly indicates less discrepancy between the *z'* values of *S-D* and Wheeler's factor than for either *F* or *K*. This correspondence is some joint function of shape and elevation. The greatest gap between *S-D* and *F* and *K* in relation to the Psychotic factor is on the psychotic scales of *Pa, Pt,* and *Sc*. Thus, *S-D* better represents the factor than either *F* or *K* where the factor plays its biggest role.

These findings would seem to indicate that much of the approach to the MMPI which *F* and *K* attempt to characterize can better be accounted for by a single factor. This underlying factor is better estimated by the *S-D* scale than by either *F* or *K*. Such an interpretation is in line with the general findings of the studies cited above in which it has been noted that most of the variance on objective-type personality tests can be accounted for in terms of social desirability.

The obvious heavy saturation of several of the clinical scales with what Wheeler terms a Psychotic factor but which here would seem to be as well labeled as social-*un*desirability, suggests that mental illness, in a generalized sense of the term, may be characterized as involving behavior which is socially disapproved. This is indicated over a fairly wide range of symptom-descriptive items, but particularly on the psychotic scales of *Pa, Pt,* and *Sc*.

Summary

A scale of social desirability (*S-D*) was constructed from the MMPI, based on unanimous judgments of 10 judges. It was hypothesized that this *S-D* scale would correlate significantly with the *F* and *K* validity scales of the MMPI and also with the nine clinical scales. Both hypotheses were confirmed.

The *S-D* scale correlated highly with *F* and *K* and higher with each of them than they do with each other. This suggests that a common factor underlies the three and that *S-D* is a better estimate of this factor. It was concluded that test-taking attitudes toward the MMPI can be characterized as readiness or lack of readiness to respond to socially desirable items.

The *S-D* scale had a very highly significant average correlation with the nine clinical scales. The correlations were generally higher but the pattern very similar in shape to the pattern between *F* and *K* and the clinical scales, again implying a common underlying factor. Findings from a factor analysis done elsewhere were compared with the data from this study. It was demonstrated that the distribution of loadings among the clinical scales of a Psychotic factor was quite similar to the correlation profile with the clinical scales of *S-D, F,* and *K*. *S-D* was a better estimate of this Psychotic factor than either *F* or *K*. It was concluded that a common factor underlies many of the clinical scales, particularly *Pa, Pt,* and *Sc,* and that it may be characterized as social desirability.

The findings of this study support the evidence from other studies cited that much of the variance of objective-type personality inventories can be accounted for in terms of a dimension of social desirability. The findings also appear to reflect society's well-known negative attitudes toward mental illness of labeling it as socially undesirable behavior.

Received August 23, 1955.

References

1. Edwards, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *J. appl. Psychol.,* 1953, **37**, 90–93.
2. Edwards, A. L., & Horst, P. Social desirability s a variable in Q technique studies. *Educ. psychol. Measmt,* 1953, **13**, 620–625.
3. Edwards, A. L. *Personal Preference Schedule, test manual.* New York: Psychological Corp., 1954.
4. Fordyce, W. E. Applications of a scale of dependency to concepts of self, ideal-self, mother and father. Unpublished doctor's dissertation, Univer. of Washington, 1953.
5. Hathaway, S. R., & McKinley, J. C. A Multiphasic Personality Schedule (Minnesota): I. Construction of the schedule. *J. Psychol.,* 1940, **10**, 249–254.
6. Navran, L., & Stauffacher, J. C. Social desirability as a factor in Edwards Personality Preference Schedule performance. *J. consult. Psychol.,* 1954, **18**, 442.
7. Taylor, Janet A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.,* 1953, **48**, 285–290.
8. Wheeler, W. M., Lehner, G. F., & Little, K. B. The internal structure of the MMPI. *J. consult. Psychol.,* 1951, **15**, 134–141.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

The Effect of Manifest Anxiety on a Concept Formation Task, a Nondirected Learning Task, and on Timed and Untimed Intelligence Tests

Aron W. Siegman

Veterans Administration Hospital, Bronx, N. Y.[1]
and Teachers College, Columbia University

Incidental to a validational study of the Taylor personality scale of manifest anxiety (MAS) (9), data were obtained about the following manifest anxiety correlates which have been discussed in the literature.

1. Goldstein has stated that anxiety interferes with abstract thinking (1). A recent study by Wesley (12) failed to support this hypothesis. However, the author as well as reviewers (6) have pointed out that the results of her study are by no means conclusive.
2. Recent research suggests that stress and anxiety interfere with nondirected or incidental learning (10, 11).
3. A number of investigations indicate that time pressure has a greater disruptive effect on anxious than on nonanxious *S*s (2, 8).

Procedure

Thirty-five male medical and psychiatric patients were administered the Taylor MAS (7), the Wechsler Adult Intelligence Scale (WAIS), Raven's Progressive Matrices (PM), the Bender Gestalt (BG) and the BG Recall Test (5). All *S*s were at least of normal intelligence and none was psychotic or suspected of having cortical damage.

In order to determine the effect of manifest anxiety on abstraction, *S*s' MAS scores were correlated with their scores on PM, which involves the eduction of patterns and relations.

As a test of nondirected learning the Recall phase of the BG was used. In the learning phase, usually perceived as a drawing ability test, *S*s were instructed to copy nine designs, which were presented one at a time. Following this phase of the test, *S*s were asked to reproduce as many designs as they could recall. For experimental control purposes the same test was administered to a group of 33 similar *S*s, with directions to memorize the designs as they were copying them because they would be asked to recall them. Time spent on the learning phase was the same for the two groups. The MAS scores of both groups were correlated with their scores on the Recall Test.

Finally, in order to determine the effect of time pressure on anxious and nonanxious *S*s, the WAIS was divided into two parts: time-limited subtests (arithmetic, digit symbol, picture completion, picture arrangement, block design, and object assembly) and subtests with no time limit (all the remaining subtests). The performance of the 10 highest Taylor scorers (35 and higher) and of the 10 lowest Taylor scorers (15 and less) on these two sets of subtests were compared.

Results

The correlation between MAS scores and PM scores was $r = -.41$, which for 34 *df* is significant beyond the .01 confidence level.

For the nondirected learning group the correlation between MAS scores and BG Recall

---

[1] From the Clinical Psychology Section of the Neuropsychiatric Service at the Veterans Administration Hospital, Bronx, N. Y. The author wishes to express his appreciation to Dr. H. L. Flowers, Chief of Neuropsychiatry, for his interest and support and to Dr. R. S. Morrow, Chief Clinical Psychologist, and Dr. Julia C. Hall, Assistant Chief Clinical Psychologist in charge of research, for their encouragement and help.

Table 1

Mean Scores on Timed and Untimed WAIS Subtests in Relation to Taylor MAS Scores

Group	*N*	Mean timed tests	*SD*	Mean untimed tests	*SD*	*t*†
High Taylor Scorers	10	10.10	1.81	11.48	2.15	2.51*
Low Taylor Scorers	10	11.71	2.12	10.86	2.03	1.24

* Significant at the .05 confidence level.
† Fisher's *t* for correlated means.

Test scores was $r = -.44$, which is significant beyond the .01 confidence level. However, for the directed learning group the correlation was $r = .07$ which is clearly not significantly different from zero.

Finally, Table 1 indicates that only the high MAS scorers had significantly lower scores on the time-limited subtests of the WAIS.

Discussion

The results of this study seem self-explanatory. However, the difference between the findings of the present study and those of Wesley (12) invites discussion. Whereas in the present study a negative correlation between MAS scores and PM scores was obtained, in the Wesley study the high MAS scorers did consistently though not significantly better on the Wisconsin Card Sorting Test (WCST), also a test of abstraction. That anxiety is not always a disruptive force and actually in some cases may be a motivating and facilitating force, has been pointed out by philosophers and psychologists (3, pp. 226–234). This differential effect of anxiety leads to the obvious but very important question: when does anxiety act as a disruptive force, and when does it act as a motivating and facilitating factor? One possible reply is that it depends on the degree of anxiety. Up to a certain point anxiety may be a facilitating force, but beyond this optimum point anxiety becomes a disruptive force. It is quite probable that the *S*s of the present study, being hospitalized patients, had more than the optimum level of anxiety. Also, whereas in the Wesley study the correct response was reinforced, no such reinforcement took place in the PM. Lack of reinforcement, especially on a task which was part of a diagnostic evaluation, may have increased *S*s' anxiety beyond the optimum level.

Recent research suggests that the effect of anxiety upon performance depends upon the difficulty of the task involved or, more precisely, upon the initial relative strengths of correct and incorrect response tendencies in the situations (4, 12). Thus the difference between Wesley's and the present findings may be due to the fact that the WCST is a much simpler task than the PM, and therefore less subject to disruption by anxiety. Moreover, the facilitating effect of anxiety in the WCST may be due to the fact that during the learning phase of the WCST the correct response is reinforced which increases the relative strength of the correct response.

Preliminary research suggests that motivation to do well is another important variable in determining whether anxiety will act as a disruptive or a facilitating force. The hospital files contained 14 high Taylor MAS scorers (35 and higher) whose average standard score on anxiety-sensitive tasks (arithmetic, digit span, digit symbol, block design, and object assembly of the WAIS, BG Recall Test and PM) (9) was below their average score on the remaining subtests of the WAIS. However, the hospital files also contained 9 high Taylor scorers whose average standard score on the anxiety-sensitive tasks was above their average standard score on the tasks not sensitive to anxiety. An investigation of the behavioral descriptions in the psychological reports about these two groups of *S*s revealed that whereas a majority of the good performers on the anxiety-sensitive tasks were described as cooperative, well motivated, and interested in the tests, a majority of the poor performers on the anxiety-sensitive tasks were described as apathetic, poorly motivated, and

Table 2

Test Performance of High Taylor MAS Scorers in Relation to Motivation

Motivation as reported in diagnostic report	Superior performance on anxiety tests	Inferior performance on anxiety tests
Cooperative, interested, well motivated	6	1
Reluctant, poorly motivated, apathetic	0	10
No description of *S*s' motivation	3	3

Note.—The chi-square value of the distribution in this table is 13.10, which for 2 *df* is significant beyond the .01 confidence level. However, due to the low frequencies this probability is merely suggestive.

reluctant to take the tests. Table 2 gives a more detailed analysis of the behavioral descriptions of the two groups. These results suggest that one's need and desire to achieve is an important variable in determining whether anxiety is going to have a disruptive or a facilitating effect. Thus the difference between Wesley's and the present findings may be due to the fact that the Wesley population consisted of college students who probably were much more motivated to do well on a concept-formation task than was the population of the present study, and therefore anxiety did not act as a disruptive force.

Summary

Thirty-five psychiatric and medical patients were administered the Taylor MAS, the WAIS, the PM, the BG with the usual instructions, and the BG Recall Test. Another group of 33 similar *S*s was administered the Taylor MAS, the BG, with instructions to learn the designs, and the BG Recall Test.

The correlation between MAS scores and scores on the PM was $r = -.41$. In the nondirected learning group the correlation between MAS scores and scores on the BG Recall Test was $r = -.44$, but in the directed-learning group the correlation was $r = .07$. Finally, *S*s who received high scores on the MAS obtained significantly lower scores on the timed than on the untimed subtests of the WAIS.

These findings suggest that anxiety has a disruptive effect on abstraction, incidental learning, and timed intelligence tests.

It was suggested that the nature of the task, the degree of anxiety, and motivational factors are important variables in determining whether or not anxiety will have a disruptive effect.

Received July 19, 1955.

References

1. Goldstein, K. *The organism*. New York: American Book Co., 1939.
2. Matarazzo, J. D., Ulett, G. A., Guze, S. D., & Saslow, G. The relationship between anxiety level and several measures of intelligence. *J. consult. Psychol.*, 1954, **18**, 201–205.
3. May, R. *The meaning of anxiety*. New York: Ronald, 1950.
4. Mayzner, M. S. A critical review of recent experimental studies relating anxiety to learning. *Psychol. Newsltr*, 1954, **5**, 117–137.
5. Peak, R. M., & Quast, W. A. A scoring system for the Bender-Gestalt. Hastings, Minn. (Box 292) and Minneapolis, Minn. (2810 42nd St.): Authors, 1951.
6. Taylor, D. W., & McNemar, Olga W. Problem solving and thinking. *Annu. Rev. Psychol.*, 1955, **6**, 455–482.
7. Taylor, Janet A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.
8. Sarason, S. B., Mandler, G., & Craghill, P. G. The effects of differential instructions on anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47**, 561–565.
9. Siegman, A. W. Cognitive, affective, and psychopathological correlates of the Taylor MAS. *J. consult. Psychol.*, 1956, **20**, 137–141.
10. Siegman, A. W. Some effects of mild electric shock on undirected learning. Paper read at East. Psychol. Ass., Atlantic City, March, 1956.
11. Silverman, R. E. Anxiety and the mode of response. *J. abnorm. soc. Psychol.*, 1954, **49**, 538–542.
12. Wesley, Elizabeth L. Perseverative behavior in a concept-formation task as a function of manifest anxiety and rigidity. *J. abnorm. soc. Psychol.*, 1953, **48**, 129–134.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

The Influence of Ego-Involvement on Relations Between Authoritarianism and Intolerance of Ambiguity[1]

Anthony Davids

Brown University and Bradley Home

After repeated failure to replicate previously reported findings between ethnocentricism and rigidity (19), Brown (3) discovered that a crucial variable to be considered in this research was the condition under which the subjects (*Ss*) were assessed. He found that when the *Ss* were administered the California F scale and the Einstellung arithmetic problems in a friendly, casual, relaxed atmosphere there was no relation between authoritarianism and rigidity.[2] However, when the *Ss* were administered these same measures under formal, somewhat threatening, ego-involved conditions there was a significant positive association between authoritarianism and problem-solving rigidity. On the basis of these findings, Brown concluded, "An adequate operational definition of this rigidity must include the establishment of an ego-involving testing atmosphere. The 'same' measure of rigidity employed in a relaxed testing atmosphere produces scores which are not related to authoritarianism" (3, p. 475).

In a recent paper, Davids (6) reported finding no relation between authoritarianism and intolerance of ambiguous visual or auditory stimuli. Since the *Ss* in the study were volunteers participating in a long-term research project, were examined under exceptionally comfortable, nonthreatening conditions, and were guaranteed anonymity, it seems possible that the negative findings might be attributable to the assessment conditions. The present experiment is an attempt to discover if a highly ego-involving atmosphere will lead to significant relations between authoritarianism and ambiguity tolerance.

Specifically, there are two main purposes in this investigation: one is to see if the previously reported positive findings (6) of significant relations between authoritarianism and maladjustment can be replicated; the other is to see if the negative findings in regard to relations between authoritarianism and intolerance of ambiguity will be found when the *Ss* are highly ego-involved in the experimental tasks. In view of the import of authoritarian personality theory (1) and the controversial findings that have been reported by investigators who have worked with this theory (3, 4, 6, 9, 14, 17), the attempt to replicate previous studies and to clarify the contradictory empirical evidence seems particularly warranted.

Method

Subjects

Twenty-two male undergraduates were secured from a college employment office. They were well-matched with the 20 *Ss* used in the previous study (6) in regard to such variables as socioeconomic background, religion, academic performance, and extracurricular activities. They differed from the previous group of *S*s in that none of them were psychology majors, they had not volunteered to take psychological tests, and all were registered at the employment office as seeking employment.

---

[1] This study was carried out at the Harvard Psychological Clinic. It was supported in part by the Laboratory of Social Relations and in part by grant M-700 from the National Institute of Mental Health, Public Health Service.

[2] Since the F scale (authoritarianism) and the E scale (ethnocentricism) have been shown to correlate .77 (1) they are frequently used interchangeably and results are often generalized from one scale to the other. Also, the concepts "rigidity" and "intolerance of ambiguity" have been used interchangeably (11, 12) and both have been applied to high authoritarians. In the present report we will not attempt to differentiate between these concepts.

Experimental Treatment

A staff psychologist contacted the student employment office and said that he would like to interview approximately 20 students in order to select a few research assistants. Individual appointments were arranged, and the psychologist conducted a personal interview with each prospective employee. The interview took place in the psychologist's office and a formal, serious attitude was maintained throughout the interview. An attempt was made to uncover basic information about the person, as well as getting him ego-involved in the assessment procedure. Each *S* was told that we planned to hire one or two students to assist us with psychological research projects and that the persons selected for the positions would be paid at a considerably higher rate than the usual student pay scale. After indicating strong interest in securing the position, as all of the 22 *S*s did, the *S* was told that a battery of psychological tests would be administered and that our selection would be based on the test results. Thus, it is evident that these individuals took the experimental measures under conditions that were far from relaxed. They needed work, a good job depended upon their test performance, and the tests were administered in a very formal, competitive, business-like atmosphere. It seems justified to conclude that these men were highly ego-involved in the task at hand. Following completion of the test battery, the *S*s were paid for their time, and shortly thereafter each person was sent a personal letter informing him that owing to failure of a research grant to come through we would be unable to add any assistants to our staff.

Measures

Authoritarianism. The *S*s were administered the standard 30-item F scale (1). Total scores ranged from a low of 63 to a high of 122, and the mean score per item was 3.14.

Ego structure. On the basis of the personal interviews, the clinical psychologist rank-ordered the *S*s on "ego structure." A detailed definition of this concept has been presented by Murray and Kluckhohn (18) who indicate that good ego structure is highly correlated with good personal and social adjustment. In the present study, the psychologist did not have a great deal of information on which to base his evaluations, and it should be recognized that there is considerable room for error in this ranking. However, the psychologist knew that he would have to rank-order the *S*s on this dimension and tried to secure information and clinical impressions that he deemed pertinent to an evaluation of ego structure. At the completion of each interview, he made notes and assigned ratings on several personality variables and a rating on ego structure. He then used this material to assist him in making his final judgments of the *S*s' relative standing on "ego structure."

Manifest anxiety. The *S*s were administered the Taylor scale of manifest anxiety (20). Scores ranged from 1 to 29, with a mean of 12.

Psychosomatic inventory. The *S*s were administered this inventory which is supposed to provide a measure of neuroticism (15). High scores indicate normality and low scores indicate neuroticism. Scores for the present *S*s ranged from 37 to 364, with a mean of 254.

Academic achievement. Transcripts of the *S*s' academic records were secured from the registrar's office, and they were rank-ordered according to their grade-point average. These averages ranked from A − to D +, with a mean performance at the B − level.

Reactions to ambiguous visual stimuli. The *S*s were individually administered the inkblot concepts from the McReynold's Concept Evaluation Technique (16) as modified by Eriksen (8, 9). Detailed description of this procedure is presented elsewhere (6, 8, 16). Briefly, the procedure consists of pointing out various Rorschach concepts to the *S* and asking, "Could this be a ——?" Half of the 50 concepts are scored plus and half are scored minus according to Beck's (2) frequency tables. In the present experiment, the number of rejections of the 50 concepts (i.e., saying "No, it could not be") ranged from 11 to 35, with a mean rejection score of 23.

Reactions to ambiguous auditory stimuli. The *S*s were administered the Azzageddi Test, which is an auditory projective technique consisting of passages of spoken communication containing contradictory and irreconcilable statements and ideas (6, 7). Although each statement, by itself, is meaningful and coherent, when several statements are intermingled into a passage of speech, there is much confusion and contradiction inherent in the passage. Consequently, the *S* is confronted with confusing and contradictory ideas and asked to recall as many of the ideas as he can from each passage. The total number of phrases and statements in the test is 112. The number of items recalled by the present *S*s ranged from a low of 35 to a high of 70, with a mean recall score of 55. After hearing, and recalling, the eight passages which constitute this test, the *S*s were presented with sheets on which they could indicate their personal reactions to the auditory projective test. On 6-point rating scales, they indicated the degree of ambiguity they perceived in the spoken material, and the degree of satisfaction or dissatisfaction they experienced while attempting to cope with the demands of this auditory test.

Predictions

The predictions were the same as those tested in the earlier study (6) where a detailed discussion of the rationale and background of theoretical formulation is presented. Briefly, these predictions were derived from the theory of authoritarian personality (1) and from previous experimental findings (10, 11, 12, 13, 19). Specifically, it was predicted that authoritarianism would be positively associated with the number of rejections of Rorschach concepts, manifest anxiety, and neuroticism, and would be negatively associated with the number of ideas recalled on the auditory test, grade-point average, and ego structure. That is, we expected that *S*s who score relatively high on authoritarianism (F scale) would be: (*a*) less intelligent (low grade-point average), (*b*) more anxious and maladjusted (high Taylor score, low P-S Inventory score, low rank on ego structure), and (*c*) more intolerant of ambiguity (high rejection of ambiguous Rorschach concepts and low recall of ideas contained in ambiguous spoken communications).

Results and Discussion

The results presented in Table 1 indicate that only one prediction is confirmed. There is a significant negative correlation between scores on the F scale and intelligence as measured by academic achievement. This finding of a negative relation between authoritarianism and intelligence is in agreement with findings reported in several previous investigations (1, 5, 6, 13).

Turning to the failure to find a significant correlation between authoritarianism and measures of adjustment, it seems highly probable that the condition under which the *S*s were assessed does not permit a fair test of the predicted relation. Since these *S*s thought

Table 1

Rank-Order Intercorrelations Among Experimental Measures

($N=22$)

Measure	Ego structure	College grades	Auditory test	Rorschach test	Manifest anxiety	P-S Inventory
F scale	−.23	−.60**	+.10	+.30	+.25	−.27
Ego structure		−.14	+.22	+.22	−.36*	+.41*
College grades			+.04	−.21	−.08	−.01
Auditory test				−.06	−.01	+.02
Rorschach test					−.31	+.25
Manifest anxiety						−.67**

* Significant at the .05 level for a one-tailed test.
** Significant at the .01 level for a one-tailed test.

the test results would be used for psychological screening, they undoubtedly distorted their responses to the questionnaires designed to measure anxiety and neuroticism. Statistical test of the differences between the Taylor scale and P-S Inventory scores of the present *S*s and the scores obtained by the research *S*s in the previous study show that on both measures the mean scores in the present experiment indicate significantly ($p = .01$) less anxiety and neuroticism. In the earlier study (6), highly significant relations were found between authoritarianism, manifest anxiety, and weak ego structure. In view of the fact that the two inventory measures of personal adjustment were significantly different in the present experiment, and the fact that the present clinical evaluation of ego structure was based on less information and information gained in an interview in which the *S*s were trying to make a favorable impression, there is no doubt that the test of our hypothesized relation cannot be considered a crucial test. It is interesting to note, however, that the three measures of adjustment—ego structure, manifest anxiety, and neuroticism on the P-S Inventory—intercorrelate significantly. Thus, it appears that if the *S*s were consciously trying to create an impression of good personal and social adjustment, as we infer, they were able to present a consistent, though distorted, picture. Under this condition of high ego-involvement, with much to be gained by presentation of a mature, well-adjusted, personality picture, we did not find a significant relation between the measures of authoritarianism and the measures of personality. However, because of the peculiar nature of the assessment condition, which was necessary for testing the effect of ego-involvement on relations between authoritarianism and intolerance of ambiguity, we cannot accept the present negative findings as conclusive evidence.

For one thing, it should not be overlooked that each of the measures of adjustment correlated in the predicted direction with the measure of authoritarianism. All these correlations are in the vicinity of .25 and, though they are not statistically significant and are not of equivalent magnitude to those obtained in the earlier study, the fact that we secured even this magnitude of relation between our measures, as predicted, makes one think that there may well be a relation between authoritarianism and maladjustment. In view of the fact we could not replicate our previous findings, however, together with the series of negative findings reported by another investigator (17), it seems highly probable the relation between authoritarianism and indices of personal adjustment may be of limited generality. That is, the type of *S*s being examined, the setting in which they are examined, rapport, ego-involvement, and a host of additional variables may have important effects on this proposed relation. Very definitely, there is need for further systematic experimentation, with more conclusive findings, before it would seem justified to conclude that there is a *general* relation between authoritarianism and personality maladjustment.

At this point, let us consider the failure to confirm the prediction concerning the relation between authoritarianism and intolerance of ambiguity. The assessment condition was designed expressly to permit further test of Brown's (3) finding that under high ego-involvement there was a significant correlation between the F scale and measures of rigidity or intolerance of ambiguity. The results shown in Table 1 provide definite evidence that, in the present study, the hypothesis was not confirmed. Chi-square tests of association between authoritarianism and personal reactions to the auditory projective test support this conclusion. There was no significant relation between the F scale and ratings of degree of ambiguity perceived in the auditory stimuli ($\chi^2 = .19$) and no significant relation between the F scale and ratings of degree of liking or disliking for the auditory test ($\chi^2 = .74$). In other words, even though in this experiment the *S*s were examined under a condition of exceptionally high ego-involvement, which should be conducive to securing the predicted positive relation, we again found no significant association between authoritarianism and measures of tolerance of ambiguous visual stimuli and ambiguous auditory stimuli. Thus, in two independent studies we have been unable to secure results in keeping with predictions derived from reports of previous research.

Two points should be emphasized in regard to the present findings. The first is that the F scale scores do not appear to be consciously distorted. Contrary to the findings of significantly less manifest anxiety and neuroticism in these *S*s than in the research *S*s in the earlier study, the mean F scale score for these *S*s is actually slightly higher than the mean for the previous group. Thus, we infer that these *S*s were not consciously trying to appear less authoritarian and that their F scale scores are probably valid. The second noteworthy point is that the correlation between the F scale and number of Rorschach rejections is not far below the acceptable level of statistical significance and is considerably higher in magnitude than the comparable coefficient obtained under non-ego-involved conditions. We certainly can place little faith in nonsignificant findings, but we would like to interpret the present finding as at least suggestive that ego-involvement may be an important determinant of the hypothesized relation. At any rate, further study of this proposed determinant seems both necessary and worthwhile.

Since other investigators, employing a variety of experimental tasks, have also failed to find a significant relation between authoritarianism and intolerance of ambiguity (9, 14), we would advocate considerable caution and skepticism until the contradictory evidence has been reconciled. We do not yet know the conditions or variables that influence this relation and until we have more unequivocal experimental evidence it seems best to place qualifications on the general statement that authoritarians are rigid or intolerant of ambiguity.

In conclusion, we would like to emphasize the need for replication studies in this area of research. In fact, more general practice of replication of psychological experiments would seem to be highly desirable. It appears that few investigators bother to repeat their own experiments and few have the desire to spend their research time and energy doing an experiment that has already been done by someone else. However, when exact replications or slightly modified extensions of previous studies result in similar positive findings one could then feel much more confident of the generality and consistency of the phenomena. And when several independent studies, which should logically bear upon common phenomena, yield negative or inconsistent findings the need for further experimentation and restricted generalization would become apparent.

Summary

The present study was concerned with relations between authoritarianism, intelligence, personal adjustment, and intolerance of ambiguity. This experiment is an extension of an earlier study in which *S*s who were relatively high on authoritarianism were found to be relatively low on intelligence, high on manifest anxiety, and low on ego structure. Contrary to expectation, however, in the previous study no relation was found between authoritarianism and intolerance of ambiguous visual and auditory stimuli. Since recent experimental evidence suggested that ego-involvement may be an important determinant of the relation between authoritarianism and rigidity or intolerance of ambiguity, the present investigation was designed to see if positive findings would be obtained when the *S*s were examined under highly ego-involving conditions.

Twenty-two male undergraduates underwent a battery of tests and personnel assessment procedures under the assumption that the results would be used in selecting someone for a highly desirable job. These *S*s were registered at the student employment office, were in need of work, and were eager to pass the psychological screening. Thus, it seems warranted to conclude that the assessment conditions were seen as formal, somewhat threatening, and highly ego-involving. Included in the assessment battery were a personal interview with a clinical psychologist, the F scale, the manifest anxiety scale, the P-S Inventory, Rorschach concept evaluation technique, and an auditory test consisting of ambiguous spoken communications. Also, grade-point averages were obtained from the *S*s' academic transcripts.

The only prediction confirmed in this study was the finding of significant negative corre-

lation between authoritarianism and intelligence as measured by academic achievement. Authoritarianism correlated in the predicted direction with the three measures of adjustment (ego structure, manifest anxiety, and neuroticism), but none of these coefficients were statistically significant. However, since the two inventory measures of adjustment indicate significantly less anxiety and neuroticism in the present *S*s in comparison with previously studied *S*s, it seems that the ego-involving conditions probably led to conscious distortion of responses to these measures of adjustment. Therefore, these negative findings should not be regarded as conclusive evidence.

No significant relation was found between authoritarianism and rejection of Rorschach concepts or recall of ideas contained in ambiguous spoken passages. Supporting this negative finding, there was also no relation between authoritarianism and a direct measure of ambiguity tolerance based on the *S*s' ratings of their personal reactions to ambiguous auditory stimulus material.

It was concluded that relations between authoritarianism, personal adjustment, and ambiguity tolerance are far from completely understood. It was suggested that these relations are probably of limited generality and that attention must be paid to the factors influencing them. And the need for replication studies was emphasized.

Received August 23, 1955.

References

1. Adorno, T. W., Frenkel-Brunswik, Else, Levinson, D. J., & Sanford, R. N. *The authoritarian personality*. New York: Harper, 1950.
2. Beck, S. J. *Rorschach's test.* 1: *Basic processes.* New York: Grune & Stratton, 1944.
3. Brown, R. W. A determinant of the relationship between rigidity and authoritarianism. *J. abnorm. soc. Psychol.*, 1953, **48**, 469–476.
4. Christie, R., & Jahoda, Marie. (Eds.) *Studies in the scope and method of "the authoritarian personality."* Glencoe, Ill.: Free Press, 1954.
5. Cohn, T. S. Is the F scale indirect? *J. abnorm. soc. Psychol.*, 1952, **47**, 732.
6. Davids, A. Some personality and intellectual correlates of intolerance of ambiguity. *J. abnorm. soc. Psychol.*, 1955, **51**, 415–420.
7. Davids, A., & Murray, H. A. Preliminary appraisal of an auditory projective technique for studying personality and cognition. *Amer. J. Orthopsychiat.*, 1955, **25**, 543–554.
8. Eriksen, C. W. Psychological defenses and "ego strength" in the recall of completed and incompleted tasks. *J. abnorm. soc. Psychol.*, 1954, **49**, 45–50.
9. Eriksen, C. W., & Eisenstein, D. Personality rigidity and the Rorschach. *J. Pers.*, 1953, **21**, 386–391.
10. Frenkel-Brunswik, Else. Tolerance toward ambiguity as a personality variable. *Amer. Psychologist*, 1948, **3**, 268. (Abstract)
11. Frenkel-Brunswik, Else. Intolerance of ambiguity as an emotional and perceptual personality variable. *J. Pers.*, 1949, **18**, 108–143.
12. Frenkel-Brunswik, Else. Personality theory and perception. In R. R. Blake and G. V. Ramsey (Eds.), *Perception: an approach to personality*. New York: Ronald, 1951. Pp. 356–419.
13. Gough, H. G. Studies of social intolerance: I. Some psychological and sociological correlates of anti-Semitism. *J. soc. Psychol.*, 1951, **33**, 237–246.
14. Jones, E. E. Authoritarianism as a determinant of first-impression formation. *J. Pers.*, 1954, **23**, 107–127.
15. McFarland, R. A., & Seitz, C. P. A psychosomatic inventory. *J. appl. Psychol.*, 1938, **22**, 327–339.
16. McReynolds, P. Perception of Rorschach concepts as related to personality deviations. *J. abnorm. soc. Psychol.*, 1951, **46**, 131–141.
17. Masling, J. M. How neurotic is the authoritarian? *J. abnorm. soc. Psychol.*, 1954, **49**, 316–318.
18. Murray, H. A., & Kluckhohn, C. Outline of a conception of personality. In C. Kluckhohn, H. A. Murray, & D. M. Schneider (Eds.), *Personality in nature, society, and culture.* (2nd Ed.) New York: Knopf, 1953.
19. Rokeach, M. Generalized mental rigidity as a factor in ethnocentrism. *J. abnorm. soc. Psychol.*, 1948, **43**, 259–279.
20. Taylor, Janet A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Ideational Expression of Hostile Impulses[1]

Robert D. Wirt[2]
Stanford University

That hostility is of major importance within personality structure and feelings about hostility are essential problems in psychotherapy are beliefs widely held among contemporary theorists and clinicians. Although considerable research has been done during the past decade or so on the expression of aggression, little fundamental research has been carried out to demonstrate the relative importance of hostility or its proper place in a comprehensive picture of personality dynamics. In fact it is difficult to find more than the bald statement, in most theories, that hostility is very important and that it is closely related to psychopathology. It is with this relationship between psychopathology and hostility that the present work is concerned.

It is commonly believed that the unabashed release of aggression is due to a breakdown of normal psychic control, or to a pathological increase in hostile elements of personality. Thus our laws provide special treatment for offenders whose crimes were committed during times of "temporary insanity." Expressions such as "he must have been crazy to do a thing like that" usually have reference to destructive acts.

Freud (8, 10, 11, 12, 13) put heavy emphasis on hostility as a reaction to threat to security, though in later elaborations of psychoanalytic theory (28) he shows hostility to derive from libido in combination with forces (the repetition compulsion) of the death instinct. The early formulation is greatly expanded by Dollard and his associates (3, 4, 21), but hostility becomes quite simply an innate response to frustration. Neo-Freudian theories (8, 14, 16, 18, 19, 22, 23, 25, 27, 28) view hostile expression more clearly in terms of interpersonal relations and most of these theorists see hostility as arising as a natural reaction to frustration with its control and expression being governed by mores of the society. With a breakdown of tolerance for frustration (or "ego disorganization") the individual is expected to display increased aggression and increasingly inappropriate expression of hostile behavior. Murray (22) explicitly states that there is a direct relationship between hostility and psychopathology, neurotics being more hostile than normals; and psychotics, he asserts, are most hostile of all. Cultural anthropologists and sociologists have stressed the differences in the freedom of expression of aggression among members of differing social groups (2, 29). Presumably this leads to differences in the degree to which hostility is a problem for members of these groups.

In general, then, there is some agreement among various theorists that hostility is of central importance in psychopathology and that its expression is modified by social experience.

Several questions are raised by this summary. If hostility is directly related to the degree of psychopathology, the broadest meaning of this assertion must be true, at the very least: normal individuals will express less hostile ideation than neurotics, who will express less than psychotics. Hostile ideation must be a precursor (or concomitant) of aggressive behavior since no matter how repressed the hostility may be, if it is really a problem it must eventually, no matter in how disguised a form, be expressed ideationally. Furthermore, if the expression of aggression is bound by cultural controls one might reasonably expect to find differences in the amount of hostile ideation expressed among different sex, age, educational, and socioeconomic groups; presumably women and persons more frequently and longer exposed to the societal values have more opportunity and more pressure to learn the control of hostility and less freedom for hostile expression. Beyond this there may be differences in the mode of such expression.

---

[1] This report is based in part on a dissertation submitted to the Department of Psychology and the Committee on Graduate Study of Stanford University, April, 1952. The writer wishes to express his appreciation to Professors C. L. Winder, Maud A. Merrill, and Douglas Lawrence for their generous help in this investigation.

[2] Now at the University of Minnesota.

Procedure

Finney's Palo Alto Aggressive Content Scale (6, 7, 26) was used as the measure of hostile ideation. This scale is based upon the Rorschach content. Rorschach (24) originally suggested his technique as a test of ideation. It provides a standardized situation in which symbolic as well as less disguised thought processes are elicited. The Finney scale was used, instead of similar scales, such as that by Elizur (5) because its scoring is more objective (15); it has distinguishable categories of hostility and it has been demonstrated to separate validly groups of known assaultive individuals from nonaggressive persons (26). Four categories for classifying aggressive responses are provided: Derogatory Remarks, Victim of Destruction, Possibly Destructive, and Active Destruction. Any response can be categorized as nondestructive or as falling into one or two of the above categories of hostility.

Three groups were taken for the sample: a group of 76 normal, 32 neurotic, and 50 schizophrenic individuals. An effort was made to stratify the groups in such a way that they would compare with the general population in age, education, and socioeconomic status. The normal group contained 38 men and 38 women, selected from the community among persons without psychiatric diagnoses. All in the abnormal groups were male patients at the Palo Alto Veterans Administration Hospital. These were individuals whose diagnoses were not complicated by known organic conditions or by admixture of other psychiatric reactions and for whom the Rorschach did not contribute in the diagnostic study. The characteristics of the sample are described more fully elsewhere.[3]

Table 1

Summary Data from Distribution of Raw Scores

Distribution of responses	Normal ($N=76$)	Neurotic ($N=32$)	Schizophrenic ($N=50$)
Total number:*			
M	23.03	23.72	17.80
SD	11.76	16.20	7.88
Aggressive responses:†			
M	6.96	10.19	7.60
SD	4.31	10.74	6.20
Nonaggressive responses:			
M	16.39	14.31	11.32
SD	9.55	9.73	8.05
"Victim" responses:			
M	2.60	5.00	4.44
SD	2.17	7.84	5.16
"Aggressor" responses:‡			
M	4.36	5.19	3.16
SD	3.34	5.05	3.11

* Since an aggressive response can be scored in more than one category, the total number of responses will not be the exact sum of number of aggressive plus number of nonaggressive responses.

† Aggressive responses equals the sum of "victim" plus "aggressor" responses.

‡ Aggressor responses equals the sum of Derogatory Remarks plus Possibly Destructive plus Active Destruction responses on the Finney Scale.

The Rorschach protocol of each individual was scored by the Klopfer and Kelley system (17) and the responses rated on the Finney scale. Ratings of a subsample of 393 responses gave 95 per cent agreement between two independent judges.

Results

Table 1 contains the summary data for the distribution of aggressive responses. The "total number of responses" is *R* in Rorschach terms, and it includes all responses, both hostile and nonhostile; "number of aggressive responses" refers to the sum of all four categories in the Finney scale; "number of nonaggressive responses" includes those responses not contained in one of Finney's categories; number of "victim responses" refers to the category of Victim of Destruction responses in the scale; and number of "aggressor" responses" refers to the sum of all four categories: Derogatory Remarks, Possibly Destructive, and Active Destruction. An aggressive response can be scored in as many as two categories. In the sample used here there was an average of less than one-half response per record requiring double classification.

---

[3] Wirt, R. D. Pattern analysis of the Rorschach. To be published.

The mean number of responses for the schizophrenic group, 17.80, is considerably less than that of the neurotic group, 23.72, or of the normal group, 23.03. The correlation between number of responses and number of aggressive responses is .60 for normals, .83 for the neurotic group, and .63 for the schizophrenic group. This suggests, as one might expect, that there is a tendency for an increasing number of aggressive responses to be associated with increased responsiveness. Total number of responses tends to be associated with any measure on the Rorschach that is quantitative, since the more responses an individual gives the more likely he is to increase the number of responses in all categories. In order to partial out the influence of the number of responses, an analysis of covariance was computed with number of aggressive responses and total number of responses as variables. The same procedure was used to test the differences between pairs of groups. The analyses show demonstrable differences in the hostile ideation of these groups. Table 2 gives the results.

As the results shown in this table indicate, each group differs from each other group, and there is a highly significant difference among the three groups. When the means are adjusted according to the results of the covariance analysis by the method suggested by McNemar (20, p. 328), it can be seen (Column 4, Table 3) that the groups do not fall into the predicted order.

The control factors were tested by two methods: analysis of covariance and a chi-square method suggested by Cronbach (1, p. 411). This method successfully controls the effect of total number of responses when differences such as number of aggressive responses are being studied among groups. Distributions covering the entire range of each factor were made for three categories of age, four categories of education, and four categories of socioeconomic status of each group and between the sexes for the normal group. None of these was related to the expression of hostile ideation in significant degree. In the case of the abnormal groups the relationship was even less pronounced than in the normal group, but in no group did either the chi square or the *F* ratio approach an acceptable level of statistical significance. The Pearson correlation between age and number of aggressive responses was only − .04.

Table 2

F Ratios for Analysis of Covariance for the Three Groups

Groups	*F* Ratio
Normal vs. Neurotic vs. Schizophrenic	6.69**
Normal vs. Neurotic	8.69**
Normal vs. Schizophrenic	7.52**
Neurotic vs. Schizophrenic	5.83*

* Significant beyond the .05 level.
** Significant beyond the .01 level.

Table 3

Means of Aggressive Responses

Group	Raw-score means	Adjusted means	Rank order
Normal	6.96	6.36	1
Neurotic	10.19	9.38	3
Schizophrenic	7.60	9.07	2

Number of aggressive responses was divided into those representing "victim" and those representing "aggressor," and Cronbach's chi-square procedure applied. This method failed to show any difference in direction of punitiveness for the social classes. However a clear tendency for direction of punitiveness to be related to psychopathology was found. The *F* ratio for analysis of covariance for "victim" responses was 7.46, which for 2 and 154 degrees of freedom, is significant beyond the .001 level. The trend is for the normals to show most "aggressor" responses with the schizophrenics showing significantly more "victim" responses and the neurotic group falling in between and showing no characteristic direction.

Discussion

Clearly these results lead to the conclusion that it requires a complex analysis of personality factors in addition to hostility to classify individuals, even grossly, along a continuum of psychopathology. Individuals in all of the groups were found throughout the entire range of ideational expression in all of the categories of aggression. It would seem that hostility cannot be taken *necessarily* to be a critical factor in the mental health of an individual. But the group differences are significant and the severely ill individuals were shown to be more expressive of hostile ideation than the normal group. Furthermore, there was a definite tendency for the groups to show differences in the direction of expression of hostile impulses. This investigation suggests that those theories relating hostility to psychopathology are substantially correct, but it will require further elaboration for a theory to account for individual differences and for the greater ideational expression of hostile impulses by neurotic individuals.

Hostility certainly is a critical factor in our social intercourse. This research seems to confirm the assertions of some theorists that hostility is common, primitive, and pervasive. But few differences, and no significant ones, were found among different social, economic, age, or educational groups. This would suggest that the effectiveness of methods of socialization in the handling of these impulses leaves much to be desired. It may be that one group of individuals is as hostile as another, but that the behavioral expression of their hostility differs widely, while its ideational expression is similar. Given appropriate channels of expression the impulse probably should not be labeled pathological, unless we wish to be in the untenable position of labeling nearly everybody as pathological.

Summary

Having reviewed a number of theoretical positions it was concluded that most contemporary theorists believe hostility to be a crucial factor in the mental health of individuals and to be subject to cultural pressures. Groups of normal, neurotic, and psychotic subjects of differing age, educational, and social status (including both men and women in the normal group), were compared on the amount of ideational expression of hostile impulses found on scoring their Rorschach protocols using Finney's Palo Alto Aggressive Content Scale. A statistically significant difference in hostile ideation was found to exist among the three groups and between any two. The normals showed the least hostility, next were the schizophrenics, and the neurotic group scored the most. The groups showed a significant difference in the direction of punitiveness, with the normal group tending to direct hostility outward, the schizophrenics inward, while the neurotic group showed no characteristic direction. None of the age, educational, social, or sex differences were statistically significant in the amount or direction of hostile ideation.

The results were interpreted to mean that many contemporary theorists are in error in stating that hostility is directly related to profundity of illness and, further, they are in error in assuming that great hostility is necessarily an indication of pathology. However, it was emphasized that the results do clearly demonstrate a tendency for greater expression of hostility among groups of severely mentally ill individuals as compared with normals and that the direction of expression characteristically differs. It was pointed out that the lack of support for theories claiming that hostility is amenable to social pressure is not, per se, proof that cultural, pedagogic, efforts are limited, but rather may mean that the present methods used to control the expression of hostility are inadequate. Further research into the relationship between ideational and behavioral expression of hostile impulses is indicated.

Received August 17, 1955.

References

1. Cronbach, L. J. Statistical methods applied to Rorschach scores. *Psychol. Bull.*, 1949, **46**, 393–429.
2. Davis, A., & Havighurst, R. L. Class and color differences in child rearing. In C. Kluckhohn, & H. A. Murray (Eds.), *Personality in nature, society and culture*. New York: Knopf, 1948. Pp. 252–264.
3. Dollard, J., Miller, N. E., Doob, L. W., Sears, R. R., Ford, C. S., Hovland, C. I., & Sollenberger, R. T. *Frustration and aggression*. New Haven: Yale University Press, 1939.
4. Dollard, J., & Miller, N. E. *Personality and psychotherapy*. New York: McGraw-Hill, 1950.

5. Elizur, A. Content analysis of the Rorschach with regard to anxiety and hostility. *Rorschach Res. Exch.,* 1949, **13**, 247–285.
6. Finney, B. C. Projective test correlates of assaultive behavior. Unpublished doctor's dissertation, Univer. of California Library, 1950.
7. Finney, B. C. *Instructions for the use of the Palo Alto aggressive content scale*. Palo Alto, California: Veterans Administration Hospital, 1951 (mimeographed).
8. Freud, Anna. *The ego and mechanisms of defense*. New York: International Universities Press, 1946.
9. Freud, S. *Civilization and its discontents*. London: Hogarth, 1930.
10. Freud, S. Instincts and their viscissitudes. In *Collected papers*. Vol. IV. London: Hogarth, 1950. Pp. 60–83.
11. Freud, S. *The ego and the id*. London: Hogarth, 1927.
12. Freud, S. Mourning and melancholia. In *Collected papers*. Vol. IV. London: Hogarth, 1950. Pp. 152–170.
13. Freud, S. *The problem of anxiety*. New York: Norton, 1936.
14. Fromm, E. *Escape from freedom*. New York: Farrar & Rinehart, 1941.
15. Gorlow, L., Zimet, C. N., & Fine, H. J. The validity of anxiety and hostility Rorschach content scores among adolescents. *J. consult. Psychol.,* 1952, **16**, 73–75.
16. Horney, Karen. *New ways in psychoanalysis*. New York: Norton, 1943.
17. Klopfer, B., & Kelley, D. M. *The Rorschach technique*. Yonkers, N. Y.: World Book Co., 1946.
18. Lewin, K. *Field theory in social science*. New York: Harper, 1950.
19. May, R. *The meaning of anxiety*. New York: Ronald, 1950.
20. McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.
21. Miller, N. E., & Dollard, J. *Social learning and imitation*. New Haven: Yale Univer. Press, 1941.
22. Murray, H. A. *Exploration in personality*. New York: Oxford Univer. Press, 1938.
23. Rogers, C. *Client-centered therapy*. Boston: Houghton Mifflin, 1951.
24. Rorschach, H. *Psychodiagnostics* (4th Ed.) Berne: Hans Huber, 1942.
25. Rosenzweig, S. An outline of frustration theory. In J. McV. Hunt (Ed.), *Personality and the behavior disorders*. New York: Ronald, 1944. Pp. 370–388.
26. Storment, Carlyn T., & Finney, B. C. Projection and behavior: a Rorschach study of assaultive mental hospital patients. *J. proj. Tech.,* 1953, **17**, 349–360.
27. Sullivan, H. S. *Conceptions of modern psychiatry*. Washington: The William Alanson White Psychiatric Foundation, 1947.
28. Thompson, Clara. *Psychoanalysis: evolution and development*. New York: Hermitage House, 1950.
29. Warner, L., & Lunt, P. D. *The social life of a modern community*. New Haven: Yale Univer. Press, 1941.
30. Wirt, R. D. Ideational expression of hostile impulses. Unpublished doctor's dissertation, Stanford Univer. Library, 1952.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

The Attitudes of the Mothers of Male Catatonic and Paranoid Schizophrenics toward Child Behavior[1]

Arnold P. Goldstein and Arthur C. Carr
Creedmoor State Hospital

The study was designed to determine if the attitudes of mothers of catatonic schizophrenics to child behavior differ from the attitudes of the mothers of paranoid schizophrenics, testing the hypothesis derived from the recent suggestion of Arieti (1) that catatonics are persons who as children were criticized for their actions, while paranoids were criticized for their intentions and for lying.

A questionnaire of 56 items related to child-raising attitudes was composed, borrowing from statements which Mark (2) found highly differentiating between mothers of schizophrenics and nonschizophrenics. Items from his scale and newly composed items were selected in terms of their relevance for the hypothesis.

The attitude scale was administered to 60 mothers, 34 of whom had sons who were presently diagnosed as catatonic and 26 who had sons diagnosed as paranoid. Only clear-cut cases of each disorder were chosen. The two groups' responses to these items were compared, item by item, by means of chi square.

Results of the statistical analysis offered no support for the hypothesis, as only 3 of the 56 statements differentiated the groups at the .05 level of confidence or better, a number not sufficiently in excess of that expected purely on a chance basis.

An unexpected finding, however, appeared to offer some support for the contention that catatonic schizophrenics' mothers may have a conflict over "choice of action" to a degree not found in the mothers of patients who were paranoid. Of the mothers solicited, mothers of catatonics with a significantly greater frequency reported themselves unable to complete the questionnaire, and refused to do so. Of the mothers cooperating in the study, refusal to answer individual items as reflected in the omission of answers occurred 46 times in the responses of the catatonics' mothers as opposed to only 10 times in the responses of the mothers of the paranoid patients.

The limitations of such evidence as offered by studies comparable to this cannot be denied in view of their failure to rule out the differential effects of the illness itself, as well as numerous other more complicated relationships which are difficult if not impossible to control. Evidence related to the etiology of schizophrenia, however, is at the present time *totally* so far from being unequivocal that it would appear that one can only maintain an open mind, accepting and evaluating accumulating evidence in the light of its limitations, until such time as the definitive, crucial evidence is available.

Brief Report.
Received January 19, 1956.

References

1. Arieti, S. *Interpretation of schizophrenia.* New York: Robert Brunner, 1955.
2. Mark, J. C. The attitudes of the mothers of male schizophrenics toward child behavior. *J. abnorm. soc. Psychol.*, 1952, **48**, 185–189.

---

[1] An extended report of this study may be obtained without charge from Arthur C. Carr, Creedmoor State Hospital, Queens Village, New York, or for a fee from the American Documentation Institute. Order Document No. 4814 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

The Effectiveness of Psychotherapy with Individuals Who Have Severe Homosexual Problems

Albert Ellis
New York, New York

Although early writers on homosexuality stressed its inherited character and minimized its treatability, contemporary investigators—including Allen (1, 2), Cory (3), Creadick (4), Ellis (5, 6), Anna Freud (7), Henry (8, 9), Laidlaw (12), Lewinsky (13), London and Caprio (14), Nedoma (15), Poe (16), Srnec and Freund (17), and Westwood (18)—tend to view homosexuals as suitable, if difficult, subjects for psychotherapy. Unfortunately, however, the recent literature contains no account of the treatment of a sizable number of cases, with an attempt to deal statistically with some of the important variables concerned in such treatment.

The present paper reports on 40 individuals (28 males and 12 females) who were seen for five or more psychotherapeutic sessions in the author's private practice during the years 1951 to 1955. Thirty-six of these patients were having overt homosexual relations; while four (2 males and 2 females) were not overt homosexuals but were obsessed with homosexual thoughts and were afraid that they soon would become inverts.

In seeing these individuals with severe homosexual problems, an active form of psychoanalytically-oriented psychotherapy was employed, and one of the main therapeutic goals was to help the patient overcome his fear of heterosexual relations and, through improved sex-love relations with members of the other sex, to minimize his homosexual interests and activities. The therapeutic goal was *not* that of inducing the patient to forego *all* homosexual interests because, as the writer has pointed out previously (5, 6), that would be unrealistic. The neurotic element in homosexuality is not the homosexual activity or desire itself, since man is biologically a bisexual or plurisexual animal who, to some degree, may be considered rare or abnormal if he has absolutely *no* homoerotic desires or participations during his entire lifetime. The abnormality in homosexuality consists of the *exclusiveness*, the *fear*, the *fetishistic fixation*, or the *obsessive-compulsiveness* which is so often its concomitant. The aim of psychotherapy, therefore, should be to remove these elements: to free the confirmed homosexual of his underlying fear of or antagonism toward heterosexual relations, and to enable him to have satisfying sex-love involvements with members of the other sex.

The 40 individuals surveyed in this paper, then, were treated for their homosexual *problems* or *neurosis* rather than for their homosexual desire or activity per se. They were deemed to be distinctly or considerably improved when, during the course of psychotherapy, they began to lose their fears of the other sex, to enjoy heteroamative relations, to be effective partners in these relationships, and to lose their obsessive thoughts about or compulsive actions concerning homosexuality.

The patients seen were largely young people: 18 being under 25 years of age; 19 between 26 and 35; and 3 over 36 years. Thirty-one of the patients were single; 5 married; 4 divorced or separated. Twenty-eight were moderately or distinctly emotionally disturbed; 12 were very severely emotionally disturbed. One had a grade-school education; ten were on the high-school level; 23 had some college training; and 6 had done graduate work.

Results

The data gathered by treating 28 males and 12 females with serious homosexual problems with active psychoanalytically-oriented psychotherapy are shown in Table 1. From this table some interesting sex differences may first be noted. The female patients were more improved as a result of treatment, more often married or divorced, and more often evaluated as very severely emotionally disturbed. When tested by chi-square analysis, these differences all proved to be significant at the .05 level of confidence. There were also non-significant tendencies for the female patients, when compared to the males, to be less educationally advanced, to have had more heterosexual activity, and to be more desirous of making a better heterosexual adjustment.

The only other relationship in Table 1 that proved to be significant when tested by chi-square analysis was that between the patients' desires to achieve heterosexual adjustment and their benefiting during the period of psychotherapy. Thus, of the 10 individuals who showed little or no improvement, 5 displayed little or no desire to achieve heterosexual adjustment; and of the 30 who showed distinct or considerable improvement, 2 displayed little or no desire to achieve heterosexual adjustment. Using Yates's correction for small cell frequencies, this difference proves significant at the .01 level of confidence.

Relationship differences which are suggested by the data of Table 1 but which do *not* prove to be statistically significant include these: (*a*) The better-educated males seemed to improve more while being treated. (*b*) The less-disturbed males were more improved, but the more-disturbed females were more

Table 1

Relationships Between Benefits Received with Psychotherapy by 28 Male and 12 Female Homosexual Patients and Selected Other Variables

	Number of patients benefited with psychotherapy							
	Little or no improvement		Distinct improvement		Considerable improvement		Total	
Variable	M	F	M	F	M	F	M	F
Age								
Under 28	4	0	3	3	5	5	12	8
28 and over	6	0	4	1	6	3	16	4
Marital status								
Single	9	0	7	2	9	4	25	6
Married, divorced, or separated	1	0	0	2	2	4	3	6
Education								
Grade or high school	4	0	0	1	1	3	5	4
College or graduate school	6	0	7	3	10	5	23	8
Clinical evaluation								
Moderately or distinctly emotionally disturbed	6	0	5	2	9	2	20	4
Severely emotionally disturbed	4	0	2	2	2	6	8	8
Extent of heterosexual activity before therapy								
Little or none	8	0	7	3	5	3	20	6
Moderate or considerable	2	0	0	1	6	5	8	6
Desire to achieve heterosexual adjustment								
Little or none	5	0	0	1	0	1	5	2
Moderate	5	0	4	2	1	1	10	3
Considerable	0	0	3	1	10	6	13	7
Length of treatment								
5 to 19 sessions	8	0	4	2	5	2	17	4
20 to 220 sessions	2	0	3	2	6	6	11	8

improved with treatment. (*c*) Individuals who had engaged in little or no heterosexual activity prior to treatment were more improved during psychotherapy. (*d*) The greater the length of the treatment, the more was the tendency of the patients to improve in the course of it.

On the whole, the results—or at least the concomitants—of employing psychotherapy with individuals having severe homosexual problems seemed to be favorable. Of the 28 male patients, 10 (36 per cent) appeared to be little or not at all improved; 7 (25 per cent) distinctly improved; and 11 (39 per cent) considerably improved. Of the 12 female patients, 4 (33⅓ per cent) appeared to be distinctly improved and 8 (67⅔ per cent) considerably improved. Of the 20 male and female patients who entered therapy with little or moderate desire to overcome their homosexual problems (but who came, instead, mainly to work on other problems or to relieve their guilt over being homosexual), 10 (50 per cent) achieved some improvement and 3 (15 per cent) achieved considerable improvement in their heterosexual relations. Of the 20 patients who came to therapy with a serious desire to overcome their homosexual problems, all made some improvement and 16 (80 per cent) made considerable improvement in their sex-love relations with members of the other sex.

Discussion

One must be duly sceptical of the results of this investigation, as with the results of most other studies of psychotherapy, because the criterion of improvement utilized is dependent on the subjective statements of the patients and judgments of the investigator. Moreover, no control group of similar subjects who had *not* been therapeutically seen was utilized. Hence, all that is known, at best, is that the patients of this study improved *with*, and not necessarily *because of*, psychotherapy.

Assuming that the therapy did, to some extent, help these patients make a better heterosexual adjustment, it may be wondered why the females, in spite of their greater disturbance, improved significantly more often than did the males. There may be several possible reasons for this difference: (*a*) Through relating to a male therapist, they may have more easily overcome their fear of and/or antagonism to males. (*b*) Because they can consummate coital relations with little difficulty once they decide to risk doing so, females may be more willing than males to make a sincere effort to engage in sex-love involvements with the other sex. (*c*) Since the females in the present sample had had more heterosexual activity than had the males at the time they entered psychotherapy, they may have found it easier to give themselves more wholeheartedly to heteroerotic relationships. (*d*) The females studied were more desirous of achieving better adjustments with members of the other sex than were the males—and this desirousness was found, in the present study, to be significantly correlated with clinical improvement.

Two interrelated questions may be asked: (*a*) Why were so many more homosexual males than females seen as patients? (*b*) Of the females who were seen, why did they tend to be more disturbed than were the males?

The answer to the first question probably is that, in our society, there are many more confirmed homosexual males than females—a point on which Cory (3), Kinsey and his associates (10, 11), and other authorities seem to agree. This may be because the sexually (and generally) inadequate male in our culture will tend to be more rejected by women than the inadequate female will be rejected by men. Women, moreover, tend to seek romantic and marital heterosexual attachments in spite of their lack of physical satisfaction, or even in spite of their fear of or hostility toward men. Where many inadequate males, therefore, will resort to homosexual activity, many equally inadequate females will take refuge in frigidity.

If this is true, it may be predicted that women who do become Lesbians will tend to be more emotionally disturbed, on the whole, than male homosexuals. This is exactly what has been found in the present study. At the same time, it should be noted that 29 per cent of the males and 67 per cent of the females studied were found to be very severely emotionally disturbed—a quite high percentage in both instances. This is probably because confirmed homosexuals are, on the whole, exceptionally disturbed individuals—which one would expect to be the case with persons who overtly surrender such an important characteristic as their fundamental sex role and who, in addition, are usually severely frowned upon and persecuted by the rest of their society.

The obtained significant relationship between the patients' improvement in heterosexual relationships and their desire for such improvement seems understandable enough. This is often true for all kinds of disturbed individuals: those who really want to get better, and who will do the hard work required in the course of active psychoanalytically-oriented therapy, almost invariably improve considerably, and often in a relatively short length of time. Homosexuals, in particular, who are essentially phobic in that they fear sex-love involvements with the other sex and take what seems to be the "easier" way out of homoeroticism, can, like most other phobics, overcome their neurosis if they will (*a*) acquire insight into the origin of their fears and (*b*) begin to participate in the activity they fear. If, with the help of good motivation, they will permit a competent therapist to help, persuade, cajole, and goad them into acquiring this insight and doing what they illogically fear, they will usually succeed in overcoming their homosexual neurosis.

Assuming that the psychotherapy received by the individuals studied actually helped them in overcoming their severe homosexual problems, the writer's impression is that the following therapeutic techniques were important in this respect: (*a*) The therapist was quite accepting and noncritical in relation to the patients' homosexual desires and acts in themselves, but at the same time insistent on unmasking the neurotic motivations behind exclusive, fetishistic, and obsessive-compulsive homosexuality. (*b*) The therapist did not insist that the patients overcome *all* their homosexual tendencies, but accepted many of these tendencies as normal or idiosyncratic. He emphasized the patients' becoming more heterosexual rather than less homosexual. (*c*) The therapist showed, by his manner and verbalizations, that he himself was favorably prejudiced toward heterosexual relationships. (*d*) Special attention was usually concentrated on the patients' general antisexual attitudes and an active attack was made on his feelings of sexual guilt and shame. (*e*) In every instance there was as much focusing on the individual's general feelings of inadequacy as on his sex problems. A major goal of therapy was always the achievement of general ego-strengthening, on the assumption that exclusive homosexuality often follows from, and is in turn the further cause of, severe feelings of worthlessness. (*f*) Wherever possible, the patients were persuaded to engage in sex-love relationships with members of the other sex and to keep reporting back to the therapist for specific discussion of and possible aid with these love relationships.

Summary and Conclusions

Twenty-eight male and twelve female individuals with severe homosexual problems were seen for from 5 to 220 sessions of active psychoanalytic psychotherapy. In terms of their achieving satisfactory sex-love relations with members of the other sex, it was found that 36 per cent of the male patients were little or not at all improved; 25 per cent distinctly improved; and 39 per cent considerably improved. Of the female patients, 33⅓ per cent were distinctly improved; 66⅔ per cent considerably improved.

The female patients were found to be significantly more improved while being treated, more often married or divorced, and more often evaluated as very severely disturbed individuals than were the male patients.

A highly significant relationship was found between both male and female patients' desires to achieve heterosexual adjustment and their benefiting during the period of psychotherapy.

Although due caution must be maintained in generalizing from this study, it is felt that there are some grounds for believing that the majority of homosexuals who are seriously concerned about their condition and willing to work to improve it may, in the course of active psychoanalytically-oriented psychotherapy, be distinctly helped to achieve a more satisfactory heterosexual orientation.

Received August 23, 1955.

References

1. Allen, C. *The sexual perversions and abnormalities.* London: Oxford Medical Publications, 1950.
2. Allen, C. On the cure of homosexuality. *Int. J. Sexology,* 1952, **5**, 148–150.
3. Cory, D. W. *The homosexual in America.* New York: Greenberg, 1951.
4. Creadick, R. N. Management of the intersexuals. *South. med. J.,* 1953, **46**, 455–460.
5. Ellis, A. On the cure of homosexuality. *Int. J. Sexology,* 1952, **5**, 135–138.
6. Ellis, A. Are homosexuals necessarily neurotic? *One,* 1955, **3** (4), 8–12.
7. Freud, Anna. Clinical observations on the treatment of manifest male homosexuality. *Psychoanal. Quart.,* 1951, **20**, 337–338.
8. Henry, G. W. Pastoral counseling for homosexuals. *Pastoral Psychol.,* 1951, **2** (18), 33–39.
9. Henry, G. W. *All the sexes.* New York: Rinehart, 1955.
10. Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. *Sexual behavior in the human male.* Philadelphia: Saunders, 1948.
11. Kinsey, A. C., Pomeroy, W. B., Martin, C. E., & Gebhard, P. H. *Sexual behavior in the human female.* Philadelphia: Saunders, 1954.
12. Laidlaw, R. W. A clinical approach to homosexuality. *Marr. Fam. Living,* 1952, **14**, 39–46.
13. Lewinsky, H. Features from a case of homosexuality. *Psychoanal. Quart.,* 1952, **21**, 344–354.
14. London, L. S., & Caprio, F. S. *Sexual deviations.* Washington: Linacre Press, 1950.
15. Nedoma, K. Homosexuality in sexological practice. *Int. J. Sexology,* 1951, **4**, 219–224.
16. Poe, J. S. The successful treatment of a forty-year-old passive homosexual. *Psychoanal. Rev.,* 1952, **29**, 23–30.
17. Srnec, Dr., & Freund, Dr. Treatment of male homosexuality through conditioning. *Int. J. Sexology,* 1953, **7**, 92–93.
18. Westwood, G. *Society and the homosexual.* New York: Dutton, 1953.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Homosexuality and the Rorschach[1]

Carl J. Nitsche, J. Franklin Robinson,
Children's Service Center, Wilkes-Barre, Pennsylvania

and Edward T. Parsons
The Pennsylvania State University

This study was an attempt to determine whether certain Rorschach content categories occur more often in protocols of homosexuals than in those of nonhomosexuals.

The *S*s, 38 white males referred to the Adult Mental Health Clinic, were divided into two groups: 19 recently convicted homosexuals and 19 patients with no known homosexual activity.

Twelve content categories were selected from studies by Due and Wright (1), Ulett (4), Chapman and Reese (3), and Fein (2): mythical distortion, qualifications toward the abnormal, dehumanization, double identification, uncertainty, evasiveness, projection of feminized behavior, preoccupation with feminine apparel, castration and phallic symbolism, sexual and anatomical responses, esoteric language and artistic references, dislike of female genital symbols.

The number of responses in each of these categories was determined for each *S* of the two groups and expressed as a percentage of the total number of responses of each *S*. This was done for each of the 12 categories separately and for the 12 categories combined. The means, standard deviations, and standard error of the difference between means were computed.

The homosexual group had more responses in 9 of the 12 categories than did the nonhomosexual group, but the differences were not statistically significant at the .05 level. Mean for the homosexual group was 4.2 responses in the selected categories, and for the nonhomosexual group, 2.7. Of the selected categories, castration and phallic symbolism, and sexual and anatomical responses, occurred most often in both groups. The homosexual group had no double identification responses, and the nonhomosexual group, no double identification, evasiveness, or dislike of female genital symbols.

Brief Report.
Received February 14, 1956.

References

1. Due, F. O., & Wright, M. E. The use of content analysis in Rorschach interpretation. Differential characteristics of male homosexuals. *Rorschach Res. Exch.*, 1945, **9**, 169–177.
2. Fein, Leah G. Rorschach signs of homosexuality in forty-three male college students. *J. clin. Psychol.*, 1950, **6**, 248–253.
3. Chapman, A. H., & Reese, D. C. Homosexual signs in Rorschach of early schizophrenics. *J. clin. Psychol.*, 1953, **9**, 30–32.
4. Ulett, G. *Rorschach introductory manual.* St. Louis: Educational Publishers, 1950.

---

[1] An extended report may be obtained without charge from C. J. Nitsche, Children's Service Center, 335 S. Franklin St., Wilkes-Barre, Pa., or for a fee from the American Documentation Institute. To obtain it from the latter, order Document No. 4811 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Relationship of Sexual Adjustment to Certain Sexual Characteristics of Human Figure Drawings

Carl N. Sipprelle and Clifford H. Swensen
University of Tennessee

Sexual aspects of human figure drawings are routinely used as indices of the sexual identification and adjustment of subjects. For example, Machover (4) states that "some degree of sexual inversion was contained in records of all individuals who drew the opposite sex first in response to the instruction, 'draw a person'."

It was the object of this investigation to attempt to determine what relationship existed between the sexual adjustment of psychotherapy patients and some of the sexual aspects of their human figure drawings.

Procedure

The subjects (*Ss*) consisted of 25 men and 24 women in psychotherapy at the University of Tennessee Psychological Service Center. Two of these *Ss* were seen only for diagnostic testing and interview, 1 was seen for 4 therapy interviews, 11 were seen for from 8 to 20 therapy interviews, and the remaining 35 *Ss* were seen for more than 20 interviews. Twenty-four of the *Ss* were seen in over 50 interviews. Three of the *Ss* were in therapy with graduate students in the last year of clinical training, and the remainder were seen by Ph.D. clinical psychologists with three or more years of experience in psychotherapy.

All *Ss* were routinely given the Draw-A-Person Test when they applied to the Service Center for treatment. The sex of the first-drawn figure was noted and the drawings were rated on sexual differentiation using a scale described in a previous paper (6). The male figures for the men and women were ranked for masculinity and the female figures were ranked for femininity, by two independent judges, and the rankings for the two judges were averaged. The reliability of rankings was determined by correlating the rankings of Judge 1 with the rankings of Judge 2. The correlations for the four rankings ranged from .50 to .84, with the average falling at .63. Although the lowest reliability coefficient (.50) is not particularly impressive, it was significant at the .01 level.

The therapists rated their patients on a "sexual adjustment rating scale." This scale consisted of eight subscales, with five points on each subscale. The subscales were: (*a*) reported frequency of intercourse; (*b*) social behavior with opposite sex at dances, dates, parties, etc.; (*c*) attitude toward sexual partner; (*d*) guilt over techniques for sexual expression; (*e*) homosexual vs. heterosexual orientation; (*f*) attitude toward opposite sex; (*g*) symptoms relating directly to sex; (*h*) discussion of sex in interview. Each point on the subscale was given a numerical value, with 1 being "good" adjustment and 5 being "poor" adjustment. The ratings for the subscales were summed to obtain a total score on sexual adjustment. The arrangement of the subscales was randomized in order to minimize the possibility of halo effect. The raters were asked to rerate each of the *Ss* two months after the original ratings were made in order to check the reliability of the ratings. The reliability coefficients were all significant at well over the .01 level, and averaged .78.

The relationships obtaining between the sexual adjustment scales and the drawings were determined by the chi-square technique for each sex group. That is the men *Ss* ($N = 25$) made up one group and the women *Ss* ($N = 24$) made up a separate group.

Table 1

Relationships Among Sexual Adjustment Scales and Human Figure Drawings

Sex adjustment	Sex first-drawn figure		Sexual differentiation		Masculinity male figure		Femininity female figure	
Scales	M	W	M	W	M	W	M	W
A	3.23	2.94	.71	2.77	.65	.68	.65	.68
B	.81	1.08	.70	.75	2.61	3.00	.00	.00
C	.81	4.00*	.70	.76	.65	.00	.65	.00
D	.80	.00	2.80	.00	.00	2.71	.65	2.71
E	.90	.89	.83	2.71	.00	.68	.00	2.71
F	3.23	.89	.00	.68	.69	.68	.71	.68
G	.84	.00	3.07	2.95	.00	.72	2.80	2.87
H	3.23	.88	.00	2.71	.65	.68	.65	.68
Total	.81	.00	.00	.75	.00	3.00	.00	3.90*

* Significant at .05 level.

Results and Discussion

Of 72 computed statistics, only two were significant at the .05 level. These results are contained in Table 1. They show that women who tend to draw the female figure first also tend to be dissatisfied with their sexual partners and that the women who draw the more feminine female figure tend to be better adjusted sexually, as measured by the total score on the sexual adjustment scale. However, such a small number of significant statistics out of such a large number of computed statistics are undoubtedly due to chance. It would appear to the authors that these results are in agreement with those of Fisher and Fisher (2). Although they concluded that women who draw female figures of average femininity tend to adjust better sexually than women who draw figures that are extremely feminine or masculine, it will be noted that out of the Fishers' 48 computed statistics only eight were significant at the .05 level.

It will be noted that there was no significant relationship between the sex of the first drawn person and sexual adjustment. These findings are in agreement with those of other investigators (1, 3, 5).

These rather disappointing findings with the DAP suggest that the DAP is either an insensitive instrument or an unreliable one which is useful only as a gross indicator of psychodynamics.

Summary

Forty-nine psychotherapy patients were rated by their therapists on sexual adjustment. Several sexual aspects of the patient's human figure drawings were rated. It was found that there was no significant relationship between the patient's sexual adjustment and the sexual characteristics of their human figure drawings.

Received August 29, 1955.

References

1. Barker, A. J., Mathis, J. K., & Powers, Clair. Drawing characteristics of male homosexuals. *J. clin. Psychol.*, 1953, **9**, 185–188.
2. Fisher, S., & Fisher, Rhoda. Style of sexual adjustment in disturbed women and its expression in figure drawings. *J. Psychol.*, 1952, **34**, 169–179.
3. Hammer, E. F. Relationship between diagnosis of psychosexual pathology and the sex of the first drawn person. *J. clin. Psychol.*, 1954, **10**, 168–170.
4. Machover, Karen. *Personality projection in the drawing of the human figure.* Springfield, Ill.: Charles C Thomas, 1949.
5. Mainord, Florence. A note on the use of figure drawings in the diagnosis of sexual inversion. *J. clin. Psychol.*, 1953, **9**, 188–189.
6. Swensen, C. H. Sexual differentiation on the Draw-A-Person Test. *J. clin. Psychol.*, 1955, **11**, 37–40.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Psychological Concomitants to Rate of Recovery from Tuberculosis[1]

Louis J. Moran,[2] George W. Fairweather, Seymour Fisher, and Robert B. Morton
VA Hospital, Houston, Texas

The presence of psychological determinants in the course of recovery from tuberculosis appears to have gained wide acceptance among specialists in the treatment of tuberculosis. It is common to find in the extensive medical literature on the subject statements such as Wittkower's that, "sometimes it may be safer to assess the patient's prognosis on the basis of his personality and his emotional conflicts than on the basis of the shadow on his film" (4, p. 202).

Only in recent years, however, have attempts been made to furnish experimental evidence for this conviction, which is based almost exclusively on clinical observation. In the main, these studies take two forms. One type of study draws upon a comprehensive and intensive study of selected patients with the purpose of abstracting personality dynamics to account for differences in rate of recovery. Representative of such studies are those of Benjamin, Coleman, and Hornbein (1), and Wittkower, Durost, and Laing (4). The first investigators emphasize the status of hostility and the second investigators emphasize the nature of defenses against dependency as determinants in rate of recovery from tuberculosis.

A second type of study seeks to relate test scores to specific criterion groups of fast and slow recoverers. Typical of these are Cohen's (3) study, which used a wide variety of Rorschach scoring variables, and Brotman's (2) use of various combinations of MMPI scores. The results of these studies have been essentially negative.

The present study is designed to examine the relationship between the rate of recovery of tuberculosis patients and four psychological variables: (*a*) behavior on the ward, (*b*) attitudes toward the hospital environment, (*c*) past behavior before hospitalization, and (*d*) fantasy as measured by the Thematic Apperception Test (TAT).

Method

The length of time required for the patient to convert from positive to negative bacteriology was selected as the criterion for rate of recovery. Routine laboratory reports are made on the bacteriological status of the patient's sputum or his gastric contents. A negative report indicates the absence of tubercle bacilli in the laboratory specimen. For purposes of this study, after five successive negative laboratory reports the patient is considered to be "converted," i.e., from positive to negative, bacteriologically.

The 46 patients studied were also subjects in a national research on the effectiveness of chemotherapy. For purposes of this latter project, all subjects were excluded who presented complications for an evaluation of chemotherapy, such as previous drug treatment, certain other diseases, etc. Also, subjects whose physical condition prohibited the random assignment of certain drugs were excluded. At the follow-up four months after admission virtually no patients had converted; at the five- to eight-month follow-up, over two-thirds had converted; and at the nine- to twelve-month follow-up, virtually all patients had converted. The highly selected nature of the present sample, as well as the potent effect of chemotherapy on the criterion should be carefully considered in generalizing the results of this study to the general population of tuberculous patients.

---

[1] The authors would like to acknowledge the generous support and guidance of Drs. Daniel E. Jenkins, Hollis G. Boren, and Irving Chofnas, tuberculosis specialists, in the conduct of this study.

[2] Now at the University of Texas.

The 32 patients that converted within five to eight months were designated the fast recovery group. The 14 patients that did not convert within five to eight months were designated the slow recovery group. There was no statistically significant difference between the fast recovery group and the slow recovery group in age, race, previous bed-rest treatment, or stage of illness (i.e., minimal, moderate, far advanced). All subjects received chemotherapy during the eight-month period under study.

Measurement of behavior in the hospital.[3] The ward behavior of each patient was rated on a 64-item rating scale independently by two aides and two nurses who had usually attended the patient for several months prior to the rating. Items representing three classes of behavior were selected on a rational basis from the pool of items: (*a*) 14 items concerning conformity to regulations, e.g., staying in bed, wearing a mask, covering coughs, (*b*) 22 items concerning relations with personnel, e.g., forms of demanding, complaining, conviviality, and (*c*) 7 items concerning relations with other patients, e.g., hazing, "reporting" a patient, disturbing others.

Each item offered two alternatives, one of which was judged on an a priori basis to be adaptive. The score for a specific item was the number of times the adaptive alternative was checked. For example, if all four raters checked the adaptive alternative the score was four; if two raters checked the adaptive alternative the score was two. The patient's score on any subscale was the sum of the scores for the items comprising that subscale.

Measurement of attitudes toward the hospital environment. The attitude scale consisted of items representing three classes of attitudes, selected on a rational basis from a pool of 45 items: (*a*) 10 items concerning ward regulations, (*b*) 14 items concerning personnel, and (*c*) 4 items concerning other patients. Items were couched in terms of statements, e.g., "If I had a little more uptime I would get well just as quickly." The statement was read to the patient, who was instructed to respond in one of the following four ways: strongly agree, moderately agree, strongly disagree, moderately disagree.

In the construction of the scale, an a priori judgment was made as to whether agreement or disagreement with a particular item could be considered least or most adaptive. The response of the patient to each item was scored 0, 1, 2, or 3, depending on the intensity of the attitude and its adaptive directionality. Thus, an item considered adaptive in the direction of agreement received the following scores: strongly disagree = 0, moderately disagree = 1, moderately agree = 2, strongly agree = 3. The score for any subscale was the sum of the scores for the items comprising that subscale.

Measurement of prehospital life adjustment. This information was obtained by means of structured interviews. The total interview schedule concerning past life adjustment consisted of 132 items. From this pool of items, three rational subscales were constructed: (*a*) 7 items concerning response to institutional control, e.g., school attendance, arrests, military disciplinary actions, (*b*) 9 items concerning general interpersonal relations, e.g., fights, number of close friends, preferred activities, and (*c*) 12 items concerning past life achievement, e.g., school achievement, organizational leadership, gain in military rank, gain in occupational level.

Each possible response to an item was ranked on an a priori basis from least to most adaptive. The score for each possible response to an item was its rank position, with the least adaptive response receiving a score of zero. Thus, the response to an item with two response possibilities received a score of zero or one; an item with three response possibilities received scores of zero, one, or two, and so on through the items.

---

[3] A complete set of the measuring instruments (except for TAT), with scoring instructions, has been deposited with the American Documentation Institute. Order Document No. 4812 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for photoprints or $1.25 for microfilm. Make checks payable to Chief, Photoduplication Service, Library of Congress.

To prevent items with the greatest number of response possibilities from unduly influencing the total scale score, responses to each item were rescored on a scale from zero to nine. Thus, an item with two responses divided the zero-to-nine scale into three equal units and the two alternatives received scores of 3 and 6. A three-response item divided the zero-to-nine scale into four equal units and the three alternatives received scores of 2.3, 4.5, and 6.7, respectively. All items on the past life adjustment scale were so rescored. Any subscale score was the sum of the items comprising that scale.

Measurement of fantasy productions. Stories to ten TAT pictures were obtained from 38 of the 46 patients. The test was administered individually, with cards in the following sequence: 1, 2, 4, 12M, 8BM, 6BM, 17BM, 18BM, 6GF, 7BM. These protocols were scored by Seymour Fisher in a blind manner, i.e., without knowledge of the individual patients, of the criterion group into which any patient fell, or of the total number of cases in either of the criterion groups. Each protocol was scored in the following four ways.

1. Extent of hostile fantasy was judged by the number of instances in which a person was described as intentionally injuring, harming, deceiving, blocking or inflicting any kind of damage to another person. Hostility was scored even when the act was described as justified or as being carried out by an official agent, e.g., policeman. Hostility was not scored when the damaging act was "accidental" or due to abstract forces of nature. Interjudge rank-order reliability, based on ten independently scored protocols was .92.

2. Vague or weak vs. clear or strong depiction of parental figures was judged on the basis of stories to cards 2, 6BM, and 7BM. Parental figures were scored as clear or strong if the story described either parent in a clearly domineering or unfriendly role or in a clearly favorable or friendly role. Parental figures were scored as vague or weak if the parents were described as inadequate or if the story data were too fragmentary or unclear to permit classification in the clear or strong category. An interjudge reliability of .81 on the dichotomous scores was obtained on twenty independently scored protocols.

3. Aspiration fantasies were judged by the number of instances in which a person was described as having high aspirations, engaging in unusually hard work, attaining financial success, doing an excellent job, being obligated to great effort, or attaining any laudable aim. The judgment of aspiration fantasy included statements of determination, special purpose, and the adoption of a purposeful, challenging attitude toward the world. An interscorer rank-order reliability of .62 was obtained on ten independently scored protocols.

4. Fantasies of inactivity were judged by the incidence of all explicit references to sleeping, daydreaming or dreaming, passively giving up, being in a dazed state of consciousness, being pitifully inadequate to a task, or being dreamily preoccupied with something distant from the present situation. Lack of inactive fantasy received the highest score. Rank-order reliability of scoring, based on ten independently scored protocols was .79.

Statistical treatment. It was hypothesized that rapid improvement should be correlated positively with high scores on each of the psychological variables. This hypothesis was tested by biserial correlation except for the fantasy variables which, because of dichotomous scoring, were tested by tetrachoric correlation.

The variables were then intercorrelated and a smaller matrix derived by means of a cluster analysis of the total matrix. This derived matrix provides the cluster of variables that have a positive correlation with the criterion variable and which also are all positively correlated with each other.

It is necessary at this point to anticipate the results in order to account for the presence of an additional variable. The discovery of a significant relationship between the TAT story depiction of vague or weak parental figures and the patient's fast recovery led to speculation about the absence of parental figures in the early life of such patients. This additional *post hoc* variable, complete absence (death, divorce, or separation) of one or both parents before the patient was ten year old, was therefore included.

Results

In Table 1 is shown the correlation of each variable with the criterion of fast recovery. The probability that this distribution of correlations could have occurred by chance is $p < .01$. As indicated in Table 1, conformity to hospital regulations and good relations with personnel were significantly correlated with fast recovery, .38 and .55, respectively. Good relations with other patients also

Table 1

Correlation Between Fast Recovery and Selected Psychological Variables

Psychological variables	Correlation
Behavior toward regulations	.38**
Behavior toward personnel	.55***
Behavior toward patients	.29
Attitude toward regulations	−.11
Attitude toward personnel	.10
Attitude toward patients	−.02
Past behavior toward control	.01
Past behavior toward peers	.14
Past behavioral achievement	−.18
Fantasy hostility	.44*
Fantasy clear parent	−.61
Fantasy aspirations	.25
Fantasy (lack of) inactivity	.35
Loss of real parent	.58**

* $p < .05$ that $r \leqq .00$.
** $p < .01$ that $r \leqq .00$.
*** $p < .001$ that $r \leqq .00$.

showed a correlation of .29 with rapid improvement, which statistically was not significantly greater than zero.

Neither the attitude measures nor the measures of past behavior (prehospital) was significantly correlated with fast recovery.

The measure of fantasy hostility correlated .44, $p < .05$, with fast recovery. High fantasy aspiration and the relative absence of fantasies of inactivity were correlated .25 and .35 fast recovery, but neither correlation was significantly greater than zero. The correlation of − .61 between fantasy-clear-parent and fast recovery was in the direction opposite that predicted. This finding led to the inclusion of a *post hoc* "hunch" variable, complete absence of one or both parents before the patient was ten years old. This loss-of-parent variable was significantly correlated .58 with fast recovery.

Table 2 presents an empirically derived cluster of those variables that were positively correlated with fast recovery and were also correlated positively with each other. For convenience of the reader, fantasy-clear-parent in Table 1 has been changed to fantasy-vague-parent in Table 2, and the signs reversed.

Discussion

When the results of this study are viewed collectively, it would appear that variability in rate of recovery from tuberculosis is clearly associated with observed adaptive behavior on the ward, associated to some degree with measures of fantasy, and unrelated to expressed attitudes or to prehospital behavior, as measured.

The cluster of positive intercorrelations in Table 2 furnishes what might be considered an empirically derived "syndrome," characterizing those patients in the present sample who recovered most rapidly. A central characteristic of this syndrome might be labeled "docile amiability." The extraordinary restrictions imposed as conditions of treatment are obligingly respected; interpersonal relations with the personnel who administer these restrictions are generally very cordial; and there is some indication that relations with other patients are better than average. Hostility, when it appears, is more likely to be evident in fantasy material than in the overt behavior of the patient.

The last two variables in the cluster encourage speculation about the etiological sig-

Table 2

Derived Matrix of Positively Intercorrelated Variables

Variables	1	2	3	4	5	6	7
1. Fast recovery							
2. Behavior toward regulations	.38*						
3. Behavior toward personnel	.55**	.49**					
4. Behavior toward peers	.29	.66**	.63**				
5. Fantasy hostility	.44	.17	.04	.05			
6. Fantasy vague parent	.61*	.13	.22	.01	.19		
7. Loss of real parent	.58*	.04	.22	.48*	.57*	.35	

* $p < .05$ that $r = .00$.
** $p < .01$ that $r = .00$.

nificance of parental identification in the development of "docile amiability." Both the lack of clear representation of parental figures in the TAT stories and the absence of one or both parents during the patient's childhood are suggestive of limited identification with parental figures.

The reader is free to place any interpretation he chooses on these strictly empirical data. We entertain the hypothesis that it is the strongly identified American male that is likely to show slower recovery from tuberculosis. He is the one most likely to obstruct treatment by rebelling actively against the rigidly enforced dependency which is a condition of treatment for tuberculosis. This hypothesis receives some support from Wittkower's finding that the tuberculosis cases with poorest prognosis were those whose "lives were governed by conscious and unconscious attempts to emulate, please, or placate their ambitious, driving and, hence, ambivalently regarded parents whose image had become an integral part of their personality structure. . . . Such persons take badly to hospitalization. . . . Their pent-up tension may be discharged in general restlessness and/or undesirable activities at the hospital" (4, p. 211).

Summary

Rate of recovery from tuberculosis was estimated for 46 patients by the amount of time required for bacteriological conversion to occur. This criterion was then correlated with four classes of psychological measures: (*a*) behavior on the ward, (*b*) attitudes toward the hospital environment, (*c*) life history, and (*d*) responses to ten TAT cards.

Variability in rate of recovery from tuberculosis was found to be associated with adaptive behavior on the ward, associated to some degree with measures of fantasy, and unrelated to expressed attitudes or to prehospital behavior, as measured. A cluster of positive correlations was then derived from the total matrix, and hypotheses offered concerning personality "syndromes" associated with fast recovery and slow recovery from tuberculosis.

Received September 12, 1955.

References

1. Benjamin, J. D., Coleman, J. V., & Hornbein, Ruth. A study of personality in pulmonary tuberculosis. *Amer. J. Orthopsychiat.*, 1948, **18**, 704–707.
2. Brotman, S. An investigation of psychological and physiological factors related to rate of recovery from pulmonary tuberculosis. Unpublished doctor's dissertation, Univer. of California at Los Angeles, 1955.
3. Cohen, D. Rorschach scores, prognosis, and course of illness in pulmonary tuberculosis. *J. consult. Psychol.*, 1954, **18**, 405–408.
4. Wittkower, E. D., Durost, H. B., & Laing, W. A. R. A psychosomatic study of the course of pulmonary tuberculosis. *Amer. Rev. Tuberc.*, 1955, **71**, 201–219.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Flexor-Extensor Movement on the Rorschach[1]

Mitchell Wetherhorn
Cambridge (Minnesota) State Hospital

This study attempted to validate current interpretations of flexor and extensor movement responses on the Rorschach test. It has been suggested that extensor movement is a correlate of aggressive, masculine, dominance striving, and flexor movement is a correlate of passive, submissive, feminine striving. In order to test this hypothesis, a series of special movement plates was designed, consisting of two human figures in opposition. Special plates were needed to elicit large numbers of movement responses since the average number of responses classifiable as flexor or extensor movement on the regulation Rorschach protocol is small.

The problem, stated negatively, was that flexor-extensor movement is not a measure of ascendancy-submission nor masculinity-femininity.

The *S*s were undergraduate college students enrolled in general psychology. The group comprised 47 females and 33 males. Each *S* was administered the *Mf* scale of the MMPI, the A-S Reaction Study, and the special movement plates.

Female *S*s produced significantly more flexor movement responses on the SMP than did the male *S*s ($t = 2.45$; $p = < .05$). Males did not produce significantly more extensor responses than did the females ($t = .36$; p = n.s.). The female scores correlated as follows: A-S and flexor movement − .06, A-S and extensor movement + .10, *Mf* and flexor movement − .07, and *Mf* and extensor movement + .02. Male scores correlated as follows: A-S and flexor movement − .07, A-S and extensor movement − .19, *Mf* and flexor movement − .25, *Mf* and extensor movement + .20. None of the correlations were significant.

In conclusion, it would appear, a flexor-extensor movement continuum does not measure the same dimensions of personality as the A-S Study or the *Mf* scale of the MMPI. There is need for further validation study of flexor-extensor movement interpretations. Perhaps a projective technique such as the SMP taps a deeper level of personality than questionnaires, hence the lack of higher correlations. Nevertheless, caution should be observed in utilizing flexor-extensor movement as a measure of the variables studied.

Brief Report.
Received February 17, 1956.

---

[1] An extended report of this study may be obtained without charge from Mitchell Wetherhorn, Cambridge State Hospital, Cambridge, Minnesota, or for a fee from the American Documentation Institute. Order Document No. 4828 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

The Goal-Spurt Hypothesis and the Rorschach Test[1]

Louis J. Maradie
Dade County Board of Public Instruction, Miami, Florida

The possible relationship between color and stimulus objects and its effects upon perception have been the subject of many investigations. In discussing the influence of color upon the subject's perceptions to the stimulus cards, Roschach stated that, "There is a definite correlation between the extent of motor activity and the number of responses influenced by color perception." He apologetically added that, "The causes of this correlation remain to be discovered" (6).

Confounding the problem is the fact that although there is little experimental evidence showing that the presence or absence of color in the stimulus figures is related to the number or significance of responses, clinically we find that the color plates often do evoke significant responses indicative of the patient's affective status.

Specifically, this study is concerned with the classical assumption that the presence of color in the standard VIII, IX, and X Rorschach plates stimulates normal subjects and is responsible for the subsequent increase in the number of responses to these cards. Klopfer (3) uses the ratio of responses to Cards VIII, IX, and X divided by the total number of responses to Cards I through X inclusive. According to him, underproduction is indicated when the obtained ratio is less than .30. Overproduction is indicated when the obtained ratio goes above .40 to .50. He interprets these variables as functions of either the "stimulating" or "disturbing" effects of color.

Lazarus was among the first to subject the problem of the stimulus value of color to experimentation (4). In comparing his subjects' performance on the chromatic and achromatic inkblots, he found that the absence of color had little effect on the scoring categories when the protocols were evaluated. Using the same technique, Sappenfeld and Buker (7) found no significant differences in the number of responses evoked by Cards VIII, IX, and X which were either chromatic or achromatic.

This author (5) circumvented the artifacts which arise by retesting the same subjects under different conditions. Ten distinct systematically randomized sequential orders of standard Rorschach plates in a latin-square design were individually administered to his subjects. He found that irrespective of the particular sequential order of the cards, the position of the cards in the series was of more importance with later-appearing cards evoking more responses than earlier-appearing cards. It was not definitely shown that color did not also have some effect, since this variable was not studied by the use of achromatic test plates.

Although the above-mentioned studies (4, 5, 7) have conclusively demonstrated that an increment in productivity is not inextricably related to the presence or absence of color, they have neither stimulated a refashioning of the classical assumptions nor have they offered acceptable explanations for the increased productivity phenomenon.

Procedure

Eighty females, who were enrolled as students in two nurses' training schools in Louisville, Kentucky, volunteered to act as subjects (*S*s). They ranged in age from 17 to 35 years with a median chronological age of 18. All *S*s were Rorschach-naive and were assured that the results of their performances would be kept confidential. They were individually tested within a two-month interval.

---

[1] Part of a dissertation presented to the Faculty of the Graduate School of the University of Kentucky in candidacy for the Degree of Doctor of Philosophy. The writer wishes to express his gratitude and appreciation to Dr. Richard L. Blanton, the director of this dissertation, whose suggestions and guidance were of inestimable value. Grateful acknowledgment is also made to Dr. Graham B. Dimmick and to Dr. Hans Hahn who made available some of the test materials.

Both standard and achromatic Rorschach plates were used in this study. The achromatic figures, identical in structure and size to the standard plates, were printed in black ink only. Verlag Hans Huber of Bern, Switzerland, produced both series of figures, but the achromatic figures were mounted by the author.

The work tasks were three in number.

The Digit Symbol task is a visual-motor association task which involves the substitution of three different numbers—2, 4, and 6—under the appropriate symbol—%, #, and &. Ten 5½- by 8-inch cards, with 54 substitutions to be made on each, were presented to each of the *S*s. No symbol followed itself and each symbol was present 18 times on each card.

The Successive Addition task, a modification of Hahn's KRHH Test, is a more complex visual-motor task. Ten 5- by 7-inch cards, each composed of 5 columns of 11 one-digit numbers, were presented to each *S*. By successively adding each overlapping pair of numbers down each of the five columns, 50 sums could be calculated on each card.

The third task, Color Naming, is a modification of Stroop's Color-Word Test (9). Color Naming is a novel visual-verbal task requiring a shift from the usual reading of words to the naming of the color of the ink in which the word is printed. This stimulus material was composed of ten 6- by 8-inch cards. Each card contained five color names (Green, Red, Black, Blue, and Orange) printed incongruously in the five colors, that is, "Green" printed in blue, red, black, and orange ink, "Red" printed in green, black, blue, and orange ink. No word was printed in the color of ink which it designated. Each card contained 60 such words. No color followed itself in the same row; equal numbers of each of the colors were printed on each card.

Every card in each of the three work tasks was unique, thus minimizing the possibility of learning from trial to trial. Each card of each work task had an equal number of operations requiring approximately a total time of one minute per card to complete.

With demonstration cards identical to the actual cards of each of the three work-task cards, practice periods were given to each *S* in order to minimize any learning effects. One minute of practice on both the Color-Naming and Successive Additions tasks was given. Because of the association learning involved on the Digit-Symbol task, two minutes of practice were allowed.

In order to determine the effect of orientation toward the goal on productivity, as measured by responses on the Rorschach Test and time scores on the work tasks, two major groups were designated. Half of the *S*s composed the "Goal-Oriented" group and the remainder, the "Goal-Disguised" group.

Those in the Goal-Oriented group received instructions which indicated the actual number of operations (Rorschach cards or work-task cards) which they would be called upon to perform. The Rorschach test instructions to these *S*s were those suggested by Beck (1) in which the number of cards the *S* will receive is clearly specified. They were admonished not to make any mistakes on the work tasks by attempting to work rapidly. It was emphasized that they should work at their own normal and comfortable rate of speed.

To further insure a set for goal orientation, the stacks of ten Rorschach cards and the stacks of ten cards of each of the work tasks, appropriately numbered on the reverse side of each, were visible to the *S* and kept before him during the entire testing period.

The Goal-Disguised group was subjected to a more trying procedure in that its work task and Rorschach test instructions were completely unstructured as to orientation toward the goal. The Rorschach test instructions to these *S*s were also those suggested by Beck except for the substitution of the word *series* in place of the word *ten*. These *S*s did not know the number of cards which would be presented. The admonition concerning errors and rate of speed was identical to that given the Goal-Oriented group.

Half of the *S*s in this group were confronted with a stack of twenty Rorschach cards placed before them. Nothing was communicated to imply that they would not be expected to respond to them all. In the case of the remaining *S*s, the ten Rorschach plates were hidden from view under a box at the examiner's side and were presented one at a time.

The two disguise techniques were also carried out with alternate work tasks. With approximately half the *S*s, two of the three work tasks were hidden and the third disguised in the stack of twenty. The other *S*s were presented two of the work tasks disguised in the stack of twenty and the third task hidden at the examiner's side.

In both the Goal-Oriented and Goal-Disguised groups, half of the *S*s in each group received the standard plates and half the achromatic plates. Instead of presenting the ten cards in their usual order from I through X, a 10 by 10 latin square was constructed (5). Ten distinct sequential orders were arranged in which each card preceded and followed every other card just once, and each card appeared once in each column and once in each row. To each of these discrete orders, an *S* was randomly assigned and presented with the plates, either standard or achromatic, peculiar to that order. The same latin square was replicated seven times by the random assignment of additional *S*s to each of the ten orders of presentation. Eight *S*s responded to each of the ten dissimilar orders of presentation.

In both Goal-Oriented and Goal-Disguised groups, half the *S*s were presented the Rorschach test prior to the work tasks, while for the other half, the work tasks were followed by the Rorschach test.

Both the Rorschach test scores (number of responses per card) and the work-task scores (number of seconds per card) were treated mathematically by the analysis-of-variance technique, followed by *t* tests when appropriate.

The Rorschach Goal-Spurt ratio is composed of the number of responses to the cards in positions eight, nine, and ten divided by the total number of responses to the cards in positions one through ten.

The work-task Goal-Spurt ratio is composed of the number of seconds on cards eight, nine, and ten divided by the number of seconds on cards one through ten. A ratio was calculated for each of the three work tasks and averaged into a single ratio.

It was then determined whether an increment in a particular *S*'s productivity on the work tasks, as demonstrated by a decline in her time scores, was related to an increment in her productivity on the Rorschach, as demonstrated by an increase in the number of her responses. Each individual was ranked on the basis of the extent of her Rorschach and average work-task Goal-Spurt ratios and the correlation determined by the rank-difference method.

Results

Before combining the data from the four separate latin squares of the "Goal-Oriented" group and the four of the Goal-Disguised group, they were tested for homogeneity of total variance by means of Bartlett's test (2). The resulting chi square values of 6.16 and 5.58 were not significant and offered no evidence against the hypothesis of random sampling from a common population.

After combining the data from the separate latin squares of the Goal-Oriented group it was further ascertained that differences in productivity between the sequences (or pairs of *S*s responding to the particular order in the latin square) were not significantly different.

Table 1

Oriented Group Analysis of Variance Results

Independent observations:	*df*	Mean square	*F*
Order of presentation	9	14.74	1.35
Residual between individuals (error)	30	10.92	
Total between individuals	39		

Correlated observations:	*df*	Mean square	*F*
Cards	9	16.82	12.84**
Positions	9	21.51	16.42**
Residual from latin square	72	1.32	1.01
Residual within individuals	270	1.31	
Total within individuals	360		
Total	399		

** Significant at 1 per cent level.

Table 2

Disguised Group Analysis of Variance Results

Independent observations:	*df*	Mean square	*F*
Order of presentation	9	6.05	.46
Residual between individuals (error)	30	13.21	
Total between individuals	39		

Correlated observations:	*df*	Mean square	*F*
Cards	9	15.00	12.00**
Positions	9	2.35	1.72
Residual from latin square	72	1.44	1.15
Residual within individuals	270	1.25	
Total within individuals	360		
Total	399		

** Significant at 1 per cent level.

Table 3

Aggregate Sums of Responses to Rorschach Test Cards

	Standard Cards	Achromatic Cards	Totals
Goal Disguised	528	515	1,043
Goal Oriented	579	572	1,151
Totals	1,107	1,087	2,194

The insignificant *F* value of 1.35 in Table 1 demonstrates this fact. Similarly, there were no significant differences between sequences in the Goal-Disguised group as the *F* value of .46 in Table 2 indicates.

Table 3 reveals that total productivity in the Goal-Oriented group was slightly greater than productivity in the Goal-Disguised group although this difference did not meet statistical significance ($t = .63$). The factor of non-orientation to a goal may have accounted for the smaller number of responses.

Productivity on the achromatic plates was not significantly greater than on the standard plates ($t = .31$). This comparison is also demonstrated in Table 3 and lends credence to the belief that the presence or absence of color neither stimulates nor inhibits productivity.

The productiveness of the *S*s in the Goal-Oriented group was greater in the later sequential positions. The mean number of responses to the first seven positions was 49.4 while that of the last three positions was 76.5. The *t* value calculated for the difference between these means was 3.92, which is significant at the 1 per cent level. Testing the significance of the mean square for total responses by positions against the residual mean square within individuals, it was found that the obtained value of *F* (16.42) is highly significant for 9 and 270 degrees of freedom (Table 1).

In the Goal-Disguised group, the increment in productivity did not appear in the later sequential positions. In fact, there was a relative decrement in these later positions. The mean number of responses to the first seven positions was 53.8 while that of the last three positions was 48.2. The *t* value calculated for the difference between these means was 1.06 which does not reach statistical significance. This is also demonstrated by the insignificant *F* value (1.72) obtained between the mean square for total responses by positions against the residual mean square within individuals (Table 2).

Another major factor of interest was the effect of the cards themselves. When tested for significance against the residual mean squares, the mean squares for cards were also found to be highly significant. In the Goal-Oriented group the obtained *F* value was 12.84 (Table 1) and that of the Goal-Disguised group was 12.00 (Table 2). Both of these values were significant at the 1 per cent level for 9 and 270 degrees of freedom. Card X, in contrast to Cards VIII and IX which are also endowed with color in the standard plates, was by far the most provocative of the series. Of special importance is the fact that Card X of the achromatic plates was similarly the most provocative of this series. These findings clearly demonstrate that neither the presence or absence of color, nor the factor of goal orientation, accounts for Card X's evocative properties in this study. It is likely that the unique quality of Card X is its configuration and this is the crucial factor facilitating the *S*s responsiveness.

Whether the work tasks preceded or followed the Rorschach was not related to any differences in Rorschach productivity in either the Goal-Oriented or the Goal-Disguised group as indicated by the respective *F* values of .13 and .49.

In the Goal-Disguised group, Rorschach productivity was not affected differentially ($F = .80$) by hiding the Rorschach cards from view or presenting them in a stack of twenty.

Each of the three work tasks in the Goal-Oriented group was tested separately by the analysis-of-variance technique. Although none of the obtained *F* values approached the point of statistical significance, there were increments in productivity as the goal was approached. This increment began to occur in the ninth positions (trials) of the Digit-Symbol and Successive Additions work task and in the eighth position of the Color-Naming task.

The three work tasks in the Goal-Disguised group were also tested separately by the analysis-of-variance technique. Although none of the obtained *F* values were signifi-

Table 4

Values of t for the First Seven Versus the Last Three Positions for Rorschach and Work Tasks

	Goal oriented	p	Goal disguised	p
Rorschach	3.921	$.01 < p < .001$	1.063	$.50 < p < .10$
Digit Symbol	1.909	$.10 < p < .05$	.206	$.90 < p < .50$
Color Naming	2.233	$.10 < p < .05$	.214	$.90 < p < .50$
Successive Additions	2.216	$.10 < p < .05$	.469	$.90 < p < .50$

cant in any of the work tasks, there were decrements in productivity in the later trials of all tasks.

The values of t for the first seven vs. the last three positions for the Rorschach test and work tasks are shown in Table 4.

The Goal-Spurt ratios were determined for each *S*'s Rorschach test performance and work-task performance. They were then ranked in their order of magnitude.

In the Goal-Oriented group, the correlation between Rorschach and work-task Goal Spurts for those *S*s who received the standard Rorschach plates was .572. This was significant at the .01 level. The correlation between Goal Spurts of the *S*s who received the achromatic Rorschach plates and the work tasks was .477. This was significant at the .05 level.

In the Goal-Disguised group, the relationship between Goal Spurts was more erratic. The correlation coefficient between Goal Spurts in the standard Rorschach plates and the work tasks was − .024. For those *S*s receiving the achromatic plates and the work tasks, the coefficient was .098. Neither was significant.

Discussion

Rather than account for the increment in productivity in the all-color cards (Cards VIII, IX, and X) on the assumption of "affectivity" in response to color, the following interpretation seems more appropriate:

It is assumed that most *S*s initially interpret the Rorschach testing procedure as threatening, or at best, challenging (8). This may be a function of various factors such as the fact that the *S* is being timed, that she does not know how many or how few "things" she should see, or what "things" she should or should not see. The fact that it is a "mental" test which might expose her inadequacies may be anxiety-arousing. The fact that the large number of cards before her will involve much time and energy expenditure may in itself be resistance-arousing. As the *S* soon learns that any and all of her responses are reacted to unemotionally in a neutral, accepting atmosphere, she becomes more comfortable and the number of her responses may increase and then remain fairly constant. As she approaches the end of the series of test cards, she may feel relieved that she is about to complete the task. Along with this comes the awareness that the last cards provide the final means for demonstrating her ability as exemplified by the number of her responses. It follows that the *S*'s goal is simultaneously to produce more responses and to remove herself from the situation. The valence of the incentive is dependent upon her interpretation of the situation and the strength of her reactions to it. The extent of the increment or decrement in responses at the terminal point is, then, a reflection of the *S*'s evaluation of goal attainment and her ability to react and compensate accordingly. Since the test involves the circumvention of barriers (stimulus plates) between the *S* and the visible goal (completion of the task), the reaction to the perceived goal should have characteristics similar to that seen in learning situations (Goal Gradient Phenomena) and in work situations (End Spurt). The increment or decrement in responses can be measured by the number of responses to Cards VIII, IX, and X divided by the total number of responses to Cards I through X. This ratio is interpreted as an expression of the "Goal Spurt."

Summary

The purpose of this study was to determine the effect of knowledge of a goal upon pro-

ductivity in the Rorschach test and in a series of mental work tasks. The data were obtained from eighty Rorschach-naive student nurses. They were randomly assigned to a Goal-Oriented group, which was informed of the number of trials in each task, and a Goal-Disguised group in which no information relative to the number of trials was communicated.

Equal numbers of *S*s in both the Goal-Oriented and Goal-Disguised groups were presented with either standard or achromatic Rorschach plates. Each series of ten test plates was drawn from a predetermined latin square in which the plates were arranged into ten distinct systematically randomized sequential orders of presentation. In this manner each plate appeared in all positions and preceded and followed every other plate an equal number of times.

Three mental work tasks were also administered to each *S* in either the Goal-Oriented or Goal-Disguised condition. The ten trials (cards) of each work task provided a means of comparison between productivity in any of the ten trials (positions) of the work tasks or the ten trials (cards) of the Rorschach test.

The Goal-Spurt ratio was defined as the ratio between productivity in work-task or Rorschach test positions eight, nine, and ten divided by productivity in positions one through ten. Ratios were calculated for each individual's Rorschach performance and her average work-task performance.

Within the limits imposed by the homogeneity of the sample used, examination and statistical analysis of the data revealed that:

1. In the Goal-Oriented group, irrespective of the presence or absence of color in the Rorschach plates, the position was of major importance with later-appearing plates producing more responses than earlier-appearing ones.

2. In both the Goal-Oriented and Goal-Disguised groups, irrespective of the presence or absence of color, and independent of their orders of sequential relationships, the plates themselves differed in the number of responses elicited. Card X, because of its unique configuration, evoked more responses than any other.

3. In the Goal-Oriented group there was a significant relationship between each *S*'s Rorschach and work-task Goal-Spurt ratios. This relationship did not exist in the Goal-Disguised group.

Thus, it was shown that there is a strong positional effect present in both the Rorschach test and a selected group of visual-motor perceptual tasks. This effect is demonstrated only when the *S*s are made aware of the number of trials necessary to complete the task. In the orthodox order of presentation of the Rorschach plates, Card X, by nature of its unique configurational properties, and its final position in the series, facilitates an increment in responsiveness.

For these reasons, the "affective" ratio, or the increment in productivity on Cards VIII, IX, and X, is better explained on the basis of the Goal-Spurt formulation. This hypothesis is more parsimonious, is objectively demonstrable, is based upon studies concerned with the laws of learning and mental work and, quite apart form the various qualities attributable to color, explains the increment in productivity in the later-appearing cards.

Received September 15, 1955.

References

1. Beck, S. J. *Rorschach's test.* Vol. I. New York: Grune & Stratton, 1950.
2. Edwards, A. L. *Experimental design in psychological research.* New York: Rinehart, 1951.
3. Klopfer, B., & Kelley, D. M. *The Rorschach technique.* Yonkers, N. Y.: World Book Co., 1942.
4. Lazarus, R. S. The influence of color on the protocol of the Rorschach test. *J. Pers.*, 1948–49, **17**, 182–184.
5. Maradie, L. J. Productivity on the Rorschach as a function of order of presentation. *J. consult. Psychol.*, 1953, **17**, 32–35.
6. Rorschach, H. *Psychodiagnostics.* Berne: Hans Huber, 1942.
7. Sappenfield, B. R., & Buker, S. L. Validity of the Rorschach 8–9–10 per cent as an indicator of responsiveness to color. *J. consult. Psychol.*, 1949, **13**, 268–271.
8. Schachtel, E. G. Subjective definitions of the Rorschach test situation and their effect on test performance. *Psychiatry*, 1945, **8**, 419–448.
9. Stroop, J. R. Studies in interference in serial verbal reactions. *J. exp. Psychol.*, 1935, **18**, 643–661.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Perceptual Judgment and Associative Learning Ability of Schizophrenics and Nonpsychotics[1]

Jay L. Chambers
University of Kentucky[2]

Kasanin (5) and others have tried to analyze the thought processes of schizophrenics. Many of their observations suggested a lack of critical judgment in schizophrenics' thinking, particularly for complex problems such as planning for the future. Whether schizophrenics would also demonstrate impairment for simple judgment problems was the primary concern of this study.

A secondary aim was to determine if schizophrenic intellectual impairment also affected learning capacity. Several studies have reported little or no impairment of learning ability among schizophrenics but the controls in most of these researches were lacking or inadequate, making the results inconclusive (1, 2, 3, 4). Therefore, a testing of associative learning among schizophrenics was included in the study. A problem involving both learning and judgment was added in the eventuality that a task combining these processes would be necessary to reveal an impairment in schizophrenic thinking.

The degree of complexity of a judgment was considered, for the purposes of this investigation, to depend upon the number of variables involved in a judgment. For example, a simple judgment would involve only one variable such as the use of IQ alone in judging students' academic potential. A more complex judgment would require the combining of several variables for the judgment, as in predicting students' academic ability on the basis of the combined variables of intelligence, motivation, educational background, etc.

Perceptual judgments of spatial dimensions (e.g. length, area, etc.) were selected for the judgment problems as they afforded objective criteria, neutral or impersonal stimuli, and represented a basic type of judgment problem practiced by all cultures and individuals.

With these considerations in mind a procedure was designed to determine, if possible, whether schizophrenics, as compared with nonpsychotics, would demonstrate (*a*) impairment in complex but not in simple judgment problems, (*b*) impairment in judgment but not in learning, and (*c*) impairment in a task combining both judgment and learning but not in either ability tested separately.

Method

In order to test the various hypotheses proposed, seven tasks were devised. Each involved the sorting of geometric designs into five categories. The designs were drawn in black ink on 2¼″ × 3½″ white index cards and were presented by *E*, who held them up one at a time for *S*'s inspection and then placed them in the category chosen by *S*. In each task, the order of the designs was systematically varied so that no patterning occurred.

The perceptual judgment tasks were constructed, by preliminary experimentation with a group of normal subjects, to be of approximately equal difficulty. The difficulty of discriminating between consecutive gradations of the stimuli was held constant by maintaining a constant ratio of increment between the gradations.

---

[1] Based upon a Ph.D. dissertation performed at the University of Kentucky. The writer wishes to express appreciation to Drs. James S. Calvin, Graham B. Dimmick, Martin M. White, R. S. Allen, Ernest Meyers, and A. Dudley Roberts for assistance on the thesis. The writer is also indebted to the Lexington and Louisville Veterans Administration Hospitals, where the subjects were obtained.

[2] Now at Muskingum College.

Following are brief descriptions of the tasks, in order of their presentation.

Task 1: Associative Learning

This task required *S* to learn to sort five designs (a circle, a square, an "x," a horizontal line, and a vertical line) into five plain boxes. As there was little similarity between the designs, the task was a relatively "pure" learning problem with a minimum of discrimination required to differentiate the stimuli. Learning was by trial and error and after each choice *E* informed *S* of the correct box for the design. Each design was presented ten times in the task.

Task 2: Simple Judgment of Density of Dots

For this task *S* was required to discriminate and sort five gradations of density of dots presented on the index cards. The gradations were numbered "1" through "5" in order from the least to most dense. The use of numbers provided an already learned order of symbols. To provide an initial standard of judgment *E* showed *S* a set of the five gradations. The standard was then removed, representatives of the five gradations were presented successively by *E* to *S*, and *S* was required to judge each card for its position in the series of gradations. Each gradation was presented ten times in the task.

Task 3: Simple Judgment of Area of Circles Combined With Associative Learning

This task required *S* to judge five gradations of circles and to learn to refer to each gradation by the name of a color. (Color names for the gradations from smallest to largest were "green," "blue," "brown," "yellow," and "red.") These color referents were used for the remaining tasks. A standard of judgment was initially presented as in Task 2. Each gradation was presented ten times in the task.

Task 4: Simple Judgment of Degree of Angles

Five gradations of degree of angle were judged in this task. Task 4 did not require learning a new set of symbols for reference to the gradations, and was therefore a relatively "pure" judgment task. A standard of judgment was initially presented as in Task 2. Each gradation was presented ten times in the task.

Task 5: Simple Judgment of Length of Lines

This task was essentially the same as Task 4 except that five gradations of length of lines were used as stimuli.

Task 6: Review of Tasks 3, 4, and 5

Six trials for each gradation of circles, angles, and lines were presented to renew experience with these stimuli, because the complex judgment task to follow depended on approximately equal familiarity with each type of stimulus.

Task 7: Complex Judgment Combining Three Variables

This task was designed to test *S*'s ability to combine or integrate three variables in making a complex judgment. For each trial (63 trials), *S* was shown a card on which the three variables (a circle, an angle, and a line from the preceding judgment tasks) were all presented. The *S* was instructed to consider all three variables in making a judgment but was required to respond to the card as a unit by giving only one estimate of its position in the series of five gradations. No corrections were given on this task. For the first fifteen trials (Part a), the variables were in agreement with regard to the position in the series represented. For the remaining 48 trials (Part b), one of the cues was opposed to the other two.

The total errors constituted the accuracy score for each task. An error was defined as the difference between the position of a stimulus in the series and *S*'s judgment of its position. For Task 7b, the two cues in agreement were taken as the correct position of the stimulus.

For comparison of speed differences, the number of seconds to complete each task was recorded for each *S*.

Subjects

The *S*s were all male veterans having been in the armed forces during World War II or later. All but four of the subjects were hospitalized at either the Lexington, Kentucky or the Louisville, Kentucky Veterans Administration Hospitals. The four nonhospitalized veterans matched in other relevant respects the hospitalized subjects.

The *S*s were divided into two groups, (*a*) those psychotic at the time of testing, and (*b*) those who were not psychotic at the time of testing and who had no record of having been psychotic. All the psychotic patients carried an official diagnosis of schizophrenic reaction except one patient who later carried a mixed diagnosis including chronic brain syndrome. The nonpsychotics carried a variety of official diagnoses, the majority falling into either character disorder or neurotic categories. The two groups did not differ significantly with regard to age or intelligence. Each group was composed of 30 *S*s.

In addition to the official diagnosis, the psychotic subjects were selected on the basis of the ward psychiatrist's opinion that, de-

Table 1

Error Scores for Learning and Judgment Tasks

Type of task	Schizophrenic (*N*=30)		Nonpsychotic (*N*=30)		*t*	Level of sig.
	Mean	*SD*	Mean	*SD*		
Associative learning						
1. Symbols	54.2	22.2	55.1	20.2	.17	—
Simple judgment						
2. Dots	49.0	18.2	30.2	5.7	5.31	.001
3. Circles	56.2	18.5	45.2	12.6	2.63	.02
4. Angles	59.7	20.5	49.4	18.9	1.97	.06
5. Lines	51.9	22.6	37.3	15.0	2.89	.01
6. Review	104.3	12.6	76.1	31.0	3.00	.01
Complex judgment						
7a. {Circles Angles Lines}	16.0	7.7	10.0	4.8	3.51	.001
7b. {Circles Angles Lines}	37.0	14.5	18.7	15.7	4.59	.001

spite active psychosis, mental deterioration and incapacity was minimal. Only patients from active treatment wards were used. No patients undergoing electric shock treatment were used.

Results

The data of Task 1, presented in Table 1, indicated no significant difference between schizophrenics and nonpsychotics for accuracy of associative learning. The schizophrenics were consistently poorer in accuracy for all of the perceptual judgment tasks, however. These differences were significant at the .02 level or better, with the exception of Task 4 (simple judgment of degree of angle) where the difference attained a probability of .06.

The data presented in Table 2 indicated that the schizophrenics were slower on all the tasks. These differences were significant at the .05 level or better with the exception of Task 7, where the differences did not attain significance. A trend toward decreasing significance for speed differences, as reflected by the size of *t*, may be noted as the tasks progressed in order of presentation.

Table 2

Time Scores in Seconds for Learning and Judgment Tasks

Type of task	Schizophrenic (*N*=30)		Nonpsychotic (*N*=30)		*t*	Level of sig.
	Mean	*SD*	Mean	*SD*		
Associative learning						
1. Symbols	334.5	195.2	248.2	84.6	2.18	.05
Simple judgment						
2. Dots	326.0	185.8	212.7	35.1	3.22	.01
3. Circles	343.8	129.9	259.3	58.9	3.18	.01
4. Angles	310.5	102.4	252.5	54.4	2.69	.01
5. Lines	287.7	94.7	233.7	44.0	2.78	.01
6. Review	531.7	172.0	449.2	91.6	2.27	.05
Complex judgment						
7a. {Circles Angles Lines}	92.0	62.2	70.5	18.7	1.78	.10
7b. {Circles Angles Lines}	282.3	187.3	234.7	74.5	1.27	.30

Discussion

If accuracy is used as the criterion for learning and judgment ability, then the results of this study support the previous evidence cited from the literature. Schizophrenics showed no impairment of accuracy in learning, but were consistently less accurate than the nonpsychotics in the judgment tasks. In speed, however, the schizophrenics were slower on both learning and judgment tasks.

The finding that schizophrenics were impaired in accuracy of judgment but not in learning might be interpreted as due to differences in motivation rather than ability. Perhaps the schizophrenics were more motivated at the beginning of the testing and thus performed more accurately on the learning task than on the later judgment tasks. No consistent preferences were expressed by either group for the type of task performed but further research, varying the order of presentation of the tasks, would better test an interest or motivation influence.

The assumption that schizophrenics would demonstrate impairment for complex but not for simple judgment problems was not supported by the data. Schizophrenics were significantly poorer than nonpsychotics on both simple and complex judgment tasks. Therefore, there was no indication that degree of complexity of judgment would be a significant variable for differentiating these two groups.

An interesting aspect of the data was that schizophrenics and nonpsychotics were differentiated through the use of perceptual-type judgment problems. Clinical workers have long felt that schizophrenics revealed inadequacies in judgment of social situations and everyday life problems. Unfortunately, evaluation of such judgments is necessarily crude and unreliable. Tests of perceptual judgment of physical dimensions may be just as effective for diagnostic purposes and are more accurately and objectively evaluated.

The significant difference between the groups for Task 2, which was primarily a judgment task and required little learning, indicates that a combination of judgment and learning was not necessary to demonstrate impairment in schizophrenics. Apparently a judgment factor alone was sufficient to differentiate schizophrenics and nonpsychotics. Also the fact that the schizophrenics continued to be poorer in accuracy of judgment after Task 3, where the "color names" had become learned, offers further reason for rejecting this hypothesis.

Summary

A group of 30 well-preserved, but actively psychotic, schizophrenic subjects were compared with an equal number of nonpsychotic subjects for performance on associative learning and perceptual judgment tasks.

The results tended to confirm other research indicating impairment in schizophrenic judgment without loss of accuracy in learning. However, the schizophrenics were significantly slower on both learning and judgment tasks.

The finding that perceptual-type judgment tasks differentiated schizophrenics and nonpsychotics may be used to develop objective and accurate judgment tests for diagnostic purposes.

A combined judgment and learning task failed to disclose differences between the groups of any greater significance than was obtained by tasks which emphasized judgment alone.

Received August 23, 1955.

References

1. Bleuler, E. *Dementia praecox.* New York: International Universities Press, 1950.
2. Hall, K. R. L., & Crookes, T. G. Studies in learning impairment, I: schizophrenic and organic patients. *J. ment. Sci.*, 1951, **97**, 725–737.
3. Hull, C. L. The formation and retention of associations among the insane. *Amer. J. Psychol.*, 1917, **28**, 335–419.
4. Huston, P. E., & Shakow, D. Learning capacity in schizophrenia; with special reference to the concept of deterioration. *Amer. J. Psychiat.*, 1949, **105**, 881–888.
5. Kasanin, J. S. *Language and thought in schizophrenia.* Berkeley and Los Angeles: Univer. of California Press, 1951.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Determination of Defense Mechanisms for Conflict Areas from Verbal Material[1]

Morton Wiener, Bruce Carpenter,[2] and Janeth T. Carpenter[2]
Central State Hospital, Indianapolis

This study was undertaken to develop criteria for determining from verbal material two general classes of defense mechanisms—repressive defenses and sensitizing defenses. These criteria seem to have application to several areas of research. For example, a study on perceptual defense utilizing these criteria to select subjects has been reported by Carpenter, Wiener, and Carpenter (3). This technique might also be used to study changes in defense behavior as a consequence of therapy or for any research problem where defense mechanisms are to be determined from verbal material.

The concept of repression has been widely accepted in the literature, and there are many suggestions as to the nature of repressive mechanisms. For example, Maslow and Mittelman (8, p. 608) define repression as "the rejection and shutting out of the awareness of a reaction pattern (thought, feeling, impulse, memory) in order to avoid distress"; Shaffer (9, p. 211) defines it objectively as "a failure to make a certain response, when the stimuli are presented that might be expected to invoke it"; Thorpe (10, p. 149) defines it as "the process by means of which the individual defends himself against being confronted with unwelcome thoughts and desires." Similar definitions have been given by others (4, p. 220; **11**, p. 240).

Thus the concept of repression seems well established. The concept of sensitization as a defense mechanism has a much shorter history. Bruner and Postman (1, 2) used the term "perceptual sensitization" to describe the unexpectedly short perceptual reaction time of some *S*s to "dangerous" stimuli. They postulated a selective vigilance on the part of an organism which is characterized by meeting the "dangerous" stimulus with utmost alertness and speed. The present authors define sensitization as a heightened vigilance which operates when the environment suggests the presence of threatening material; the individual is sensitive to the conflictual stimulus but to decrease the anxiety aroused by it, he changes the significance of the stimulus (e.g., intellectualization, undoing, displacement, projection).

The primary purpose of this study is to determine whether a reliable criterion can be established for determining repressive defenses and sensitizing defenses from verbal material. Secondarily, the question of whether there is a generalized defense which individuals use for several areas of conflict or anxiety will be investigated.

Method

Subjects. One hundred forty undergraduates in a beginning psychology course made up the experimental sample.

Procedure. A special sentence-completion blank was devised for use in determining defense mechanisms. The sentence-completion blank was self-administered by each *S* outside of class. The instructions were: "Complete these sentences to express your real feelings. Try to do every one. Be sure to make a complete sentence." The sentence completion blank consisted of the following twenty stems, mimeographed with triple spacing between stems:

---

[1] This study was carried out as part of a larger project with the permission and encouragement of Clifford L. Williams, M.D., superintendent of Central State Hospital. Grateful acknowledgement is made to Dr. Charles C. Josey of Butler University for making his students available to serve as *S*s for this study. The authors wish to thank Mr. Austin Jones for serving as a judge and for critically reading a draft of this paper. Appreciation is expressed to Mr. Earl Furlow for assistance in the preliminary selection of the scoring criteria, and to the other nine psychologists who discussed the criteria in a seminar.

[2] Now at Florida State University.

1. Sports are
2. I don't want to know
3. Dancing
4. Reading is
5. I failed
6. There are
7. I resent
8. I hate
9. Sex is
10. Walking
11. Dating
12. People who neck
13. It bothers me
14. I get mad
15. I really feel
16. Men and women
17. What annoys me
18. I secretly
19. The birds
20. Deep down I

Five of the stems (Nos. 3, 9, 11, 12, and 16) were designed to elicit sexual content. Five stems (Nos. 7, 8, 13, 14, and 17) were designed to elicit content in the hostility area. Another five stems (Nos. 2, 5, 15, 18, and 20) were designed to elicit feelings about the self. The other five stems (Nos. 1, 4, 6, 10, and 19) were designed to be neutral or relatively innocuous. The stems were all relatively mild so that the *Ss* would not feel discomfort in filling out the blank. The sex and hostility stems were intended to represent several different intensities and were intended to lie along a gradient from less to more conflictual in our culture.

In scoring the completed stems another major conflict area emerged in addition to the areas of sex and hostility. Many of the stems appeared spontaneously to elicit expression of feelings of inadequacy. Since in this college group adequacy may be an important area of conflict, it was decided to score the sentence completions for sensitization to inadequacy also.

In scoring the sentence completions, the first discrimination attempted by the judges was the presence or absence of conflict in the response. Two criteria for making this discrimination were adopted: (*a*) if a defense mechanism clearly was observable in the response, it was scored as conflictual; and (*b*) if the response to the stimulus was an overreaction, underreaction, avoidance, or other inappropriate reaction to the stimulus, the response was scored as conflictual.

The experimental purpose was to dichotomize conflict responses into two broad classes of defense mechanisms, those with repressive characteristics and those with sensitization characteristics. The following criteria were used to score responses as repressive. Examples are given in parentheses:

a. The use of clichés. (*I hate* war; *Sex is* here to stay.)

b. Denial of stimulus implication. (*I hate* nobody.)

c. Avoidance of stimulus. (*I hate* asparagus.)

d. Blocking (no response).

e. Distancing from personal involvement. (*I hate* to think of how cold it is in Greenland.)

f. Very limited generalizations. (*It bothers me* when someone pops gum in my ear.)

g. Minimization of involvement in the conflictual activity. (*Sex is* something I won't know about till I'm married.)

h. Obligation, duty, imposed acceptability by authority. (*Dancing* is something young people should learn.)

i. Definitions which lead to avoidance of conflictual connotations of the stimulus. (*Dancing* is the movement of one's feet.)

j. Moralization, romanticization, naiveté, or idealization. (*Sex is* one of the gifts of God that is to be used and enjoyed as He wishes.)

The following criteria were used to score the sentence-completion responses as sensitization responses. Examples are given in parentheses:

a. Statement of inadequacy or failure. (*Sex is* something I'm very bashful about.)

b. Rationalization. (*I failed* my German exam, but it was because I was tired from working too hard.)

c. Intellectualization. (*It bothers me* to read in the newspapers that the Communists have taken over another country.)

d. Acting, then undoing or inhibiting. (*I hate* my brother, but only when he fights with me.)

e. Displacement or projection of feeling to other people. (*It bothers me* to hear people make fun of another's physical defect.)

f. Preoccupation, including elicitation of a conflictual response by a stimulus which is inappropriate for that particular content area. (*I hate* to see two girls holding hands.)

g. Projection of motives or feelings to others. (*I hate* people who deliberately hurt others.)

h. Use of humor with conflictual material. (*Sex is* something I don't know much about—I'm too innocent.)

i. Qualification. (*People who neck* are all right, if they don't do too much.)

j. Overreaction to stimulus. (*I hate* people.)

k. Denial of importance. (*Sex is* over-emphasized in our society today.)

As a general scoring criterion, avoidance of the stimulus or denial of the activity implied in the stimulus was considered a repressive mechanism; denial of the importance or conflictual nature of the activity was considered a sensitizing mechanism.

There were eight possible judgments which could be made about each response. A response could be scored: (*a*) without conflict; (*b*) hostility conflict—sensitizer; (*c*) hostility conflict—represser; (*d*) sex conflict—sensitizer; (*e*) sex conflict—represser; (*f*) adequacy conflict—sensitizer; (*g*) sensitizer response without designation of area of conflict; or (*h*) represser response without designation of area of conflict.

Four clinical psychologists acted as judges. Agreement between judges was arbitrarily defined as at least three of the four judges making the same judgment about a response from the eight possible judgments.

Each *S*'s response to each stem was typed on a separate card. All responses to each item were placed together for judging. Three judges first independently scored all 2,100 critical items. The five neutral items were not considered. Those items on which all three judges agreed were removed from the set and the remaining items were scored independently by a fourth judge. Those items which reached the agreement criterion were used to assign scores to each *S*. Items on which judges did not agree or which were judged as not indicative of conflict were not included in the scores. Each *S* was given a Represser-Sensitizer (R-S) score for sex and an R-S score for hostility. Each score was based on the number of responses in a particular conflict area judged as sensitizing and as repressive. The number of repressive and sensitizing responses were combined by a formula $(S - R + 5)$ into single score variable with scale values ranging from 0 to 11. A 0 score indicates consistent repressive defenses in an area. A score of 11 indicates consistent sensitizing defenses in an area. A score of 5 was obtained by *S*s who had an equal number of repressive and sensitizing responses in an area. Each *S* was also given a general R-S score based on the number of responses judged as repressive or sensitizing without designation of any particular area of conflict. The formula $(S - R + 5)$ was used to obtain the general R-S score. Each *S* was given an Adequacy score; this score was the number of sentences in which an expression of inadequacy was judged to have been elicited.

Results

Agreement between judges was obtained for 72 per cent of the 2,100 sentences judged. Of the remaining 28 per cent, the judges split (with two judges making one judgment and two making another judgment) on 14 per cent of the items. The remaining 14 per cent showed less agreement than this. Examination of the data indicated that certain subcategories of the scoring criteria contributed more to the unreliability of judgment than did others. For example, it was relatively easy for the judges to discriminate reliably the "denial" and "blocking" categories, but was much more difficult for the judges to differentiate reliably between "intellectualization" (sensitizing) and "distancing from personal involvement" or "limited generalization" (repressive).

To determine whether there is a tendency for individuals to use a general class of defense mechanism in different conflict areas, tabulation was made of the number of *S*s using repressive mechanisms in both the sex and hostility areas, *S*s using sensitizing mechanisms in both areas, and *S*s using sensitizing defenses in one area and repressive defenses in the other. Table 1 shows this distribution. Those *S*s having scores of 5 (indicating an equal number of repressive and sensitizing responses) in either area are not included in this table. The difference between the number of *S*s using consistent defenses and the number of *S*s using different defenses is not significant. If *S*s with scores of 5 (no

Table 1

Frequency of *S*s Using Repressive and Sensitizing Mechanisms in the Sex and Hostility Areas

	Sex	
Hostility	Represser	Sensitizer
Represser	23	32
Sensitizer	12	26

Note.—Chi square = 1.003, $p > .30$.

Table 2

Intercorrelations Among Sex R-S Scores, Hostility R-S Scores, General R-S Scores, and Adequacy Scores for 83 Subjects

	Hostility	General R-S	Adequacy
Sex	+.24*	+.32*	−.01
Hostility		+.55*	+.25*
General R-S			+.25*

* Significant at .05 level.

consistent mechanism) in one area are included in an analysis of defense by area, only 39 per cent of the 140 *S*s show consistent use of the same mechanism from one of these areas to the other.

Further investigation of the generality of defense mechanisms was made by correlating the scores obtained by 83 of the *S*s [3] for the four areas—sex, hostility, general R-S, and adequacy. Table 2 shows these correlations, which range from − .01 to + .55. All correlations except that between the sex and adequacy areas are significantly different from zero.

Discussion

The degree of agreement obtained between judges suggests that responses to the sentence-completion task can be classified fairly reliably by use of the scoring criteria. While 72 per cent agreement is far from perfect agreement, it seems adequate in view of the rather stringent agreement criterion adopted, of at least three of four judges making the same judgment from eight possible judgments.

The results of this study by and large do not support a theory of generality of defense behavior. Although there appears to be a tendency for a large minority (39 per cent) of *S*s to use the same defense mechanism for more than one area of conflict, the majority of the *S*s did not consistently use one class of defense mechanism. In addition, although significant correlations were obtained between scores representing mechanisms used in different conflict areas, these correlations are quite low. There would be little improvement over chance in attempting to predict any one *S*'s mode of defense in one area from knowledge of his defense mechanism in another area.

The findings in this study with regard to generality of defense are in agreement with the results of a study by Goldstein (5). Goldstein found that the majority of his *S*s did not exhibit consistent defense preference, although a large minority did. He concluded that there were two subgroups of *S*s, "specific defenders" and a smaller group of "general defenders."

Additional but nonquantified evidence is available which argues against a theory of consistency of defensive behavior. It became apparent in examining the scoring for individual *S*s that some *S*s use different defense mechanisms for subclasses of stimuli within a single conflict area. For example, one *S* consistently used repressive defenses in responding to heterosexual stimuli but spontaneously gave responses expressing concern about homosexuality to stimuli not expected to evoke a sexual response (by definition a sensitizing mechanism). Another *S* used sensitizing defenses in self-directed hostility responses, but used repressive defenses in other-directed hostility responses.

The lack of evidence to support a theory of generality of defenses makes rather puzzling the results of research employing *S*s selected on the basis of diagnostic labels. For example, Lazarus, Eriksen, and Fonda (7) selected *S*s diagnosed as "obsessive-compulsive" and "hysteric" personalities. When given an auditory perceptual task, the obsessive-compulsive *S*s more accurately identified threat words than did the hysteric *S*s. If defenses are specific to conflict areas rather than general, why did the hysterics consistently show repressive behavior? One possible explanation is that there is greater consistency of defense behavior among groups diagnosed as pathological than there is among "normals." Goldstein's (5) findings of greater "emotional disturbance" for the general defenders than for the specific defenders lends support to this hypothesis. However, Kurland (6) did not find significant differences between a pathological group and a normal group in response to "emotional" words on an auditory perceptual task. The question of whether pathological groups show more consistent defense behavior therefore remains unresolved.

---

[3] These 83 *S*s were those for whom data were available for several measures in a larger study. Careful examination revealed no apparent differences between those *S*s included in the correlation matrix and those not included.

Although reliability of judgment of the order found in this study is adequate for many research problems, greater reliability of judgments based on the criteria might be obtained if the subcategories of the criteria could be more sharply delineated. Experience gained through scoring the responses suggests that reliable differentiation of the subcategories depends upon response to extreme subtleties in verbal expression. As an example, to differentiate between "distancing from personal involvement" and "intellectualization," it may be necessary to respond to the intensity of the verbalization. Experience suggests that an intellectualized response reveals considerably more intensity of involvement in the verbal material than does a repressive response. A response, *"It bothers me* to read in the newspaper that the Communists have taken over another country," may be qualitatively different from a response, *"It bothers me* to read about the Communists." The specificity of the source, "the newspaper," and of the disliked activity, "taken over another country," suggests more involvement in the verbal production. As this technique is used with a variety of *S*s in a variety of situations, it may be possible to raise the reliability by delineating the scoring categories in this manner. If reliability can be increased sufficiently, use of the technique as a clinical tool might be profitably investigated.

Summary

An attempt was made to devise a technique which would permit specification of kinds of defense mechanisms used by individuals in various areas of conflict. Scoring criteria were developed to aid in identifying repressive and sensitizing responses to a sentence completion task. Judges were able to classify *S*s reliably by use of the scoring criteria. Converting judges' classifications to a single score variable allowed comparison of defense mechanisms used in several conflict areas. In general a theory of consistency of defense behavior was not supported.

Received August 29, 1955.

References

1. Bruner, J. S., & Postman, L. Emotional selectivity in perception and reaction. *J. Pers.,* 1948, **16**, 69–77.
2. Bruner, J. S., & Postman, L. Tension and tension release as organizing factors in perception. *J. Pers.,* 1947, **15**, 300–308.
3. Carpenter, B., Wiener, M., & Carpenter, Janeth T. Predictability of perceptual defense behavior. *J. abnorm. soc. Psychol.,* 1956, **52**, 380–383.
4. Dollard, J., & Miller, N. E. *Personality and psychotherapy.* New York: McGraw-Hill, 1950.
5. Goldstein, S. A projective study of psychoanalytic mechanisms of defense. Paper read at Amer. Psychol. Ass., Washington, D. C., September, 1952.
6. Kurland, S. H. The lack of generality in defense mechanisms as indicated in auditory perception. *J. abnorm. soc. Psychol.,* 1954, **49**, 173–177.
7. Lazarus, R. S., Eriksen, C. W., & Fonda, C. P. Personality dynamics and auditory perceptual recognition. *J. Pers.,* 1951, **19**, 471–482.
8. Maslow, A. H., & Mittelman, B. *Principles of abnormal psychology.* New York: Harper, 1941.
9. Shaffer, L. F. *The psychology of adjustment.* Boston: Houghton Mifflin, 1936.
10. Thorpe, L. P. *The psychology of mental health.* New York: Ronald, 1950.
11. White, R. W. *The abnormal personality.* New York: Ronald, 1948.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

The Relationship of Anxiety and "Lack of Defensiveness" to Intellectual Performance

Irwin G. Sarason
Indiana University [1]

Two studies have recently appeared in this journal dealing with the relationship between Taylor Anxiety Scale scores and ACE performance. Matarazzo *et al.* (3) found anxiety scores to be inversely related to ACE scores, while Schulz and Calvin (4) found no significant relationship between these two variables. One deficiency common to these studies was the small number of subjects employed in certain of the anxiety groups. Thus, the low anxious group of Matarazzo *et al.* included only 13 *S*s; Schulz and Calvin's high anxious group contained only 8 *S*s. One aim of the present paper is to report the results of a study similar to these but employing relatively large numbers of *S*s in all anxiety groups.

A further purpose of this study was to investigate the relation to intellectual performance of a defensive test-taking attitude as measured by the MMPI *K* scale. This seemed of interest because of the highly significant degree of relationship reported in several studies between the *K* scale and the Taylor Anxiety Scale. Pearson *r*'s of − .74 and − .81 have been reported by Heineman (1) and Matarazzo (2), respectively. Also, by including both of these scales in this study it became possible to evaluate Matarazzo's hypothesis that the Taylor scale and *K* may be used as alternate forms. If this is true, one would expect similar results for the two scales with respect to intellectual measures such as those used in this experiment.

Method

Subjects

The Biographical Inventory (5) containing the Taylor and *K* scales was group-administered to all Introductory Psychology students enrolled at Indiana University at the start of the Fall 1954 semester. The 719 *S*s used in this study include all of those students who were entering freshmen and who completed the Fall semester.

Procedure

Two measures of intellectual functioning were employed: ACE raw scores and semester over-all grade-point averages. In relating the anxiety and *K* scales to the two intellectual measures, *S*s were divided into 7 anxiety and 4 *K*-scale groups. The Taylor-scale groups were ordered from low to high anxiety at intervals of five. That is, the lowest anxiety group included *S*s with scores ranging from 1 to 5, the highest group included *S*s with scores ranging between 36 and 41. There were 29 *S*s in the lowest anxiety group and 34 *S*s in the highest group. Each of the other five groups included at least 80 *S*s.

The following are the ranges of score for the four *K* scale groups: the lowest *K* group included all *S*s with raw *K* scores between 1 and 10, the highest included *S*s with scores between 21 and 30. The remaininig two groups included *S*s with score ranges of 11–15 and 16–20. The group of low *K* scorers included 134 *S*s, the high *K* group contained 61 *S*s. The other groups each contained over 200 *S*s.

Results

Analyses of variance failed to reveal significant changes in grade-point average with increases in score on either the Anxiety scale or *K* scale. The *F* ratio for the anxiety groups was .80, while the *F* ratio for the four *K* groups was 2.20. This latter *F* was signifi-

---

[1] Now at VA Hospital, West Haven, Conn.

Table 1

Means and Standard Deviations of ACE Scores and Grade-Point Averages for *K* Scale Groupings

K-scale score	*N*	ACE Mean	ACE *SD*	G.P.A.* Mean	G.P.A.* *SD*
1–10	134	99.55	25.92	1.22	.76
11–15	296	108.23	22.92	1.35	.69
16–20	228	109.17	22.04	1.42	.70
21–30	61	107.90	23.82	1.36	.67

* Grade of C = 1, B = 2, etc.

cant between the .10 and .05 levels (3 and 715 *df*). The Pearson *r* between *K* and grade-point average was found to be .07, which was not significantly different from zero. An insignificant *F* of .83 for the mean ACE scores of the seven anxiety groups indicated a lack of relationship between increases in Taylor score and ACE total score. When an *F* was computed for the mean ACE scores of the four *K* groups, an *F* of 5.48 was obtained. With 3 and 715 *df*, this *F* is significant at the .001 level. Table 1 presents the means and standard deviations of ACE scores and grade-point averages for the 4 *K*-scale groups. Inspection of this table clearly reveals that the significance of the *F* is due to the mean of the low *K* group which differs markedly from the means of the other *K* groups. The Pearson *r* obtained between ACE and *K* scores was .09 ($p < .05$).

The *r* computed between Taylor and *K* scores was found to be equal to $-.62$ ($p < .001$). An *r* was also computed between Taylor- and *K*-scale scores for all nonfreshmen ($N = 448$) in Introductory Psychology who took the Biographical Inventory at the same time as the *Ss* in this study. The obtained correlation was $-.69$, which is significantly different ($p < .02$) from the *r* of $-.62$ for the freshmen.

Discussion

The results of this study are in agreement with Schulz and Calvin's (4) demonstration of a lack of relationship between Taylor scale and ACE scores. The finding of Matarazzo *et al.* (3) that the higher the anxiety score the lower the ACE score was in no way supported. As Schulz and Calvin suggest, one explanation for contradictory outcomes of similar studies done at different institutions may well be due to extraneous factors associated with the institutions, e.g., difference in type of student body. It is interesting that the two studies indicating no relation between Taylor and ACE scores were done at large state universities, whereas the Matarazzo *et al.* study was performed at a private and relatively smaller institution.

The results using the *K* scale are of significance in that they do not point to a gradual monotonic relationship between defensiveness scores and intellectual performance. Both with the ACE and grade-point averages, it is clear that *Ss* in the low *K* group differ markedly from *Ss* in the rest of the distribution.

The finding that low *K Ss* perform on a considerably lower level than all other *Ss* in the distribution leads, of course, to two possibilities. The obtained result could simply be due to the fact that individuals with low scores on the *K* scale are not as bright as all other individuals. Another possibility, considered more likely by this writer, is that low *K* scorers do poorly on tests like the ACE because of certain personality characteristics which are detrimental to good performance.

It may be said of low *K Ss* that they are candid and self-critical in the extreme and manifest feelings of inadequacy by too freely admitting "bad" things about themselves. In a sense they are "lacking" in defensiveness. If this is so, it is plausible that these persons when confronted with a stressful test situation in which they feel themselves being evaluated (e.g., the ACE) tend to respond with self-depreciatory attitudes. They may verbalize to themselves and recall their shortcomings (e.g., "I am dumb, I can't do it") while working at the task. To the extent that this proneness to self-criticism results in the production of responses (e.g., "Maybe I won't pass the test") unrelated to the successful completion of the task, we would expect it to be detrimental to a high level of performance. It is certainly hard to conceive of these interfering responses as being facilitative of high achievement. The result obtained in this study is what one would expect; namely, a relatively low level of performance from those *S*s who are overly self-critical.

The fact that the *F* ratio among the *K* groups was significant for the ACE and only approached significance for the grade-point averages suggests an interesting theoretical question concerning the development of an essentially intropunitive self-critical reaction to highly motivating situations in which the individual feels his ability is being evaluated. If one assumes that this lowering of the threshold for self-criticism is a learned response to situations in which the individual feels under pressure to perform well, we would expect that other possible responses more relevant to the task in question might similarly be learned. That is, the fact that someone has learned a nonadaptive habit does not preclude his eventually learning an adaptive one. For instance, individuals who feel that they are inadequate to cope with stressful situations, such as taking an intelligence test, may, if in the situation long enough and given many trials, overlearn the relevant responses for high performance sufficiently so that the detrimental effect of the learned response of self-depreciation will be kept to a minimum. With regard to the present study, it is quite possible that the low *K*'s were able to reduce the effect of their interfering self-verbalizations over the course of a semester by means of overlearning the course subject matter. On the other hand no overlearning was possible in preparation for the ACE.

A word remains to be said concerning the correlation of − .62 obtained between *K* and *A*. This correlation is considerably lower than those published in other studies (1, 2). In order to determine whether this low correlation was purely a characteristic of students at Indiana University or whether it might in part be due to the fact that all the *S*s were freshmen, the correlation between the two scales was computed for nonfreshmen as well as freshmen. The obtained correlation of − .69 for the upper-class group, while still lower than other studies, is more in line with correlations reported in the literature. It seems possible, therefore, that the two scales may become more closely associated as the length of time in college increases. We have already mentioned that the type of student body from which *S*s are sampled can be an important factor and there are certainly other socioeconomic factors which are equally relevant. At the present time it would appear that Matarazzo's suggestion (2) that the two scales be used as alternate forms is premature.

Summary

This experiment was performed to evaluate the effects of anxiety (Taylor-scale scores) and defensiveness (MMPI *K*-scale scores) on intellectual performance (ACE scores and grade-point averages). The results failed to show significant changes in these two measures of intellectual performance as a function of anxiety. However, *S*s low in defensiveness were found to perform significantly more poorly on the ACE than did all other *S*s in the *K*-scale distribution. A similar trend was found for the grade-point averages. The results were discussed in terms of the adverse effect of the self-criticism of low *K* *S*s on their performance in highly motivating situations.

Received September 12, 1955.

References

1. Heineman, C. E. A forced-choice form of the Taylor Anxiety Scale. *J. consult. Psychol.*, 1953, **17**, 447–454.
2. Matarazzo, J. D. MMPI validity scores as a function of increasing levels of anxiety. *J. consult. Psychol.*, 1955, **19**, 213–217.
3. Matarazzo, J. D., Ulett, G. A., Guze, S. B., & Saslow, G. The relationship between anxiety level and several measures of intelligence. *J. consult. Psychol.*, 1954, **18**, 201–205.
4. Schulz, R. E., & Calvin, A. D. A failure to replicate the finding of a negative correlation between manifest anxiety and ACE scores. *J. consult. Psychol.*, 1955, **19**, 223–224.
5. Taylor, Janet A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–295.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Evaluative Conceptualizations as the Basis for Clinical Judgments[1]

Franklin C. Shontz
Western Reserve University[2]

In spite of the general admiration for "objectivity" in psychology, it has frequently been observed that there is considerable room, especially in our clinical practices, for the action of "subjective," and therefore perhaps also detrimental, influences (1, 2, 3, 6). The real question is not simply whether this "subjectivity" is actually present as a factor in clinical diagnosis, but whether clinical judgments are sufficiently reliable phenomena to be considered useful as a basis for making decisions about the human personality.

Obviously, the problem is too broad to be handled in any single empirical investigation, but at least one important aspect of the clinical situation may be relatively easily approached, and shows promise for fruitful research, viz., the basis upon which psychologists make judgments regarding the "mental health" of their subjects.

We will assume for the present that, on the basis of their training, clinical psychologists are capable of building up fairly accurate descriptive conceptualizations of their patients, and that they make judgments and evaluations of these subjects on the basis of a partially subjective comparison of these descriptive conceptualizations with some standard or "ideal" picture of mental health, adjustment, or what have you. It follows, then, that two psychologists may differ in their judgments of the same individual mainly because they have grossly different "ideals" against which the subject is being compared. It also follows that it would be worthwhile if we could make these ideal conceptions explicitly measurable so as to assess their internal reliability (consistency within any single psychologist) and the extent to which clinicians would be found to agree with each other.

The purpose of the present research has been to use Q-sort method (5) in the examination of psychologists "evaluative conceptualizations" and to determine the extent to which these may be said to represent reliable and consistent diagnostic phenomena.

Method

Subjects. The subjects were twelve psychologists, all of whom were employed in the same state hospital setting. Of these twelve individuals, seven were full-time, permanent staff members and five were half-time graduate students in clinical psychology. All but one were currently engaged in diagnostic and therapeutic work with psychotic patients. The students were under the supervision of experienced staff members at all times, and none of these students had been in graduate school less than two years at the time of this investigation.

Measurements. The measuring device was a 60-item, structured-sample Q sort, administered in two equivalent forms of 30 items each. Q-sort items were selected to fill a 3 × 2 × 5 factorial scheme twice over. In this scheme, the first dimension reflected rigidity, flexibility, or disorganization of the personality processes; the second dimension reflected the direction of psychological energy either toward the self or toward others; and the third dimension defined specific item-content in terms of positive affective relationships, negative affective relationships, cognitive processes, aspirations, and techniques of adjustment. Details regarding the methods of scale construction, the means for establishing equivalent forms, and the nature of specific item-content are available elsewhere (4).

---

[1] Subproject of a dissertation presented for the degree of Doctor of Philosophy at Western Reserve University. The valuable advice and critical comments of Dr. D. W. Miles and the other members of the doctoral committee, Drs. C. S. Hall, C. F. Baker, C. R. Porter, and R. Fisher, are gratefully acknowledged.

[2] Now at Highland View Hospital, Cleveland, Ohio.

Q-sort items were sorted by the subject into nine piles to describe two separately requested conceptualizations. The 30 items of one form were first placed into the required distribution, and the task was immediately repeated with the 30 items of the equivalent form. Order of presentation of the forms was randomly distributed. The nine piles used were quasi-normally distributed in frequency of card placement from "most descriptive" to "least descriptive," and scores were assigned to each pile in the usual manner. The actual frequency distribution employed was: 1; 2; 3; 5; 8; 5; 3; 2; 1; and the scores assigned to these categories ranged from 0 for the least descriptive item to 8 for the most descriptive one.

Procedure. Each subject was asked to use both forms of the Q sort to describe his conceptualizations of a psychologically healthy person and an ideal hospitalized patient. Thus a total of 48 separate 30-item sortings were obtained.

The two frames of reference in which the cards were sorted were specified in the following set of directions:

> 1. Describe first a perfectly recovered individual, one for whom you would unhesitatingly recommend release from the hospital . . . in short, a psychologically healthy person.
>
> 2. Now describe an ideal hospitalized patient, one who might be considered for discharge if the home conditions were proper; one who would most likely get along well on the ward but who is not sufficiently healthy to warrant a recommendation for unconditional release.

The descriptions provided by the subjects were first evaluated for intrasubject reliability on the two equivalent forms of the Q sort. The mean correlations between all possible pairs of subjects were then computed to yield indices of homogeneity, or intersubject reliability on both measures.

An over-all analysis of variance was performed to determine which groups of items were significantly high or low in their descriptive scores and whether differences could be found between the two sets of descriptions obtained.

Finally, composite descriptions of each "ideal" were constructed by calculating the mean placement values (descriptive scores) for each item and by redistributing all the items back into the original quasi-normal distribution according to their average values. These composite descriptions were intercorrelated and compared in terms of item-placements as a means for specifying the similarities and differences between the two conceptualizations investigated.

Results

The mean intrasubject reliabilities of the described conceptualizations were found to be .82 for the healthy person descriptions, and .80 for the descriptions of the ideal hospitalized patient. (All figures are Pearson product-moment correlation coefficients.) These values compare satisfactorily with the reliabilities of many psychological instruments, and their magnitudes indicate that the subjects were describing conceptualizations which remained fairly stable, at least from one sorting to the next.

The mean intersubject reliabilities (homogeneities) were found to be .68 for the healthy person descriptions and .59 for the descriptions of the ideal hospitalized patient. These figures indicate a moderately high level of agreement between subjects, although they suggest that there were probably quite a few differences with respect to specific items.

The intercorrelations of the composite descriptions constructed for both "ideals" were found to be .72 on one form of the Q sort and .67 on the other. Again, there is indication that the two ideals are much the same in over-all characteristics, but there is no indication of agreement on *all* items of the measuring instrument.

Analysis of variance of the obtained sortings revealed the following significant item categories, according to the Q-sort item structure already outlined:

Items reflecting flexibility, regardless of direction or content, were found to obtain significantly high descriptive values in both conceptualizations.

Items reflecting disorganization, regardless of direction or content, were found to obtain significantly low placement values in both conceptualizations. Especially low placement was given to items reflecting disorganization of the negative affective relationships.

Items reflecting rigidity, regardless of direction or content, were found to obtain significantly low values only in the descriptions of the psychologically healthy person.

What these findings mean in terms of specific item content was determined by examining the composite sortings to locate those Q-sort statements which received both highly consistent and highly inconsistent placement when one "ideal" was compared with the other.

The following are examples of items considered by the subjects to be "desirable" (assigned to the upper extreme categories of the Q-sort distribution) in both conceptualizations: [3]

18. Considerate of others in his personal relationships.
16. Understands the feelings of others.

The following are examples of items considered by the subjects to be "undesirable" (assigned to the lower extreme categories of the Q-sort distribution) in both conceptualizations:

23. Likely to commit impulsive suicide.
57. Likely to get uncontrollably angry with others for no good reason.

The following are examples of items which were considered desirable for the healthy person, but "unimportant" (assigned to a middle category) for the ideal patient:

12. Likes people for what they are.
15. Knows his own assets and liabilities.

The following are examples of items considered to be undesirable for the healthy person and unimportant in the ideal patient:

6. Makes inflexible all-or-none judgments about people which he won't change even in the face of facts.
4. Believes in severe punishment for people he thinks are immoral.

The following item was considered desirable for the ideal patient but unimportant for the healthy person:

14. Never gets angry at people without good reason.

The following item was considered to be "undesirable" for the ideal patient but unimportant for the healthy person:

57. Takes the first opportunity he sees to satisfy an impulse.

There were no items which were considered undesirable in one description and desirable in the other.

Discussion

The present research indicates a very substantial level of over-all agreement in the evaluative conceptualizations of this group of psychologist-subjects. If we bear in mind that the reliability of the single subject in this study was about .80, it is evident that any intersubject reliability which approaches this value would indicate as great an agreement between subjects as exists within them individually. If this were the actual case, one would expect psychologists to be a sterile lot who never get into more disagreements with each other than they do with themselves on the basis of their own individual inconsistency.

Actually, these results are about what one might hope for. The figures show sufficient agreement so that we might expect this group of psychologists to be agreed at least "in principle"; but the figures do not show sufficient conformity to indicate a rigidity of conception throughout the sample. From the patient's point of view this would seem to indicate that one could be quite certain of receiving a consistent evaluation regardless of which of these psychologists were diagnosing him (as long as there were no errors in the descriptive conceptualization itself), but that specific bits of behavior might have different meaning or importance depending upon the examiner.

The "ideal" conceptualizations obtained here cannot be said to be right or wrong except on the basis of the acceptance of some uniform set of standards. The data indicate, however, that these psychologists do recognize differences between the requirements of hospital and nonhospital society. They further reveal an emphasis upon flexibility and the need for recognizing others in one's behavior. One cannot be sure what *causes* these conceptualizations, but it is apparent that this sample is highly coherent as a group. It would be interesting in the future to examine less homogeneous populations and to make comparisons between the "ideals" of behavior which characterize various professional groups, or which characterize psychologists in various cultural and social environments. It would be equally interesting to study, for example, the development of evaluative conceptualizations as a student progresses through training in psychology.

---

[3] The item numbers are those assigned in the original Q sort.

The Q-sort scale employed here was designed for use in describing the behavior of psychotic patients, and the extreme character of some of its items may have induced a level of agreement which is greater than one would normally expect. Nonetheless, this investigation has demonstrated the feasibility of examining these conceptual "ideals" by which, not only psychologists, but all people make judgments about each other.

Summary

A Q-sort investigation was made of twelve psychologists' conceptualizations of the psychologically healthy person and the ideal hospitalized patient.

Equivalent Q-sort forms were used to establish intrasubject reliabilities, and these were found to be .80 and .82 respectively for the two conceptualizations.

Average intersubject agreements were found to be .68 for the psychologically healthy person descriptions, and .59 for the descriptions of the ideal hospitalized patient. These values were interpreted as indicating a high level of general agreement between subjects, but by no means as expressive of a unity of opinion on all specific trait items.

The data were analyzed to reveal differences between the two conceptualizations, and samples of specific items were presented to illustrate the findings.

Received August 23, 1955.

References

1. Asch, S. E. Forming impressions of personality. *J. abnorm. soc. Psychol.*, 1946, **41**, 258–290.
2. Ash, P. The reliability of psychiatric diagnoses. *J. abnorm. soc. Psychol.*, 1949, **44**, 272–276.
3. Dailey, C. A. The clinician and his predictions. *J. clin. Psychol.*, 1951, **7**, 270–273.
4. Shontz, F. C. Clinical psychologists' conceptualizations of their patients. Unpublished doctor's dissertation, Western Reserve Univer., 1954.
5. Stephenson, W. *The study of behavior.* Chicago: Univer. of Chicago Press, 1953.
6. Weingartner, E. M. A study of selective perception in clinical judgments. *J. Pers.*, 1949, **17**, 369–406.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

Some Correlates of the Davis-Eells Tests

Grace T. Altus
Santa Barbara County Schools

The preponderance of research evidence indicates that there is a significant relationship between measured intelligence and socioeconomic level in that lower-class members do less well than those higher in the socioeconomic hierarchy on virtually all widely used intelligence tests. There is considerable disagreement among psychologists, sociologists, and allied social scientists as to the extent to which this positive relationship is a function of the "culture-bound" nature of the devices used to measure intelligence and the extent to which it is a function of true, innate differences in the intellectual capacities of varying occupational or social classes. Probably no one would deny that test biases are responsible for some of the measured class differences.

In the hope of minimizing, in their test, the cultural bias assumed strong in most group intelligence tests, Davis and Eells (1) in 1953 issued the Davis-Eells Test of General Intelligence or Problem-Solving Ability, otherwise known as the Davis-Eells Games. The test yields an Index of Problem-Solving Ability (IPSA) derived in the same fashion as an IQ. It is a power test, requires no reading, consists primarily of pictorial representations of multiple-choice problems for which the teacher reads the problem, and is said to be interesting and challenging to children of all social levels. It was designed for use in the first six grades of school and has a primary and elementary form.

The appeal of such a test to school personnel is strong. Many teachers, as well as laymen, are wary of standard intelligence tests. They know that not all their children are identified correctly by current tests. All of them would welcome an index of true learning ability which would identify those pupils who are and those who are not working to capacity. Is the Davis-Eells Test such an index?

Problems of test validation are nearly insurmountable if one accepts the assumption of Davis and Eells that school grades, reading skill, and other measures of academic success cannot be accepted as validation criteria since these indices *are* strongly culture-bound, whereas the whole intent of the Davis-Eells Test is to minimize such social structuring. With what, then, can one correlate the test to assess some measure of its validity? The authors seem to accept no external criteria. The validity of the test must be taken on faith—that is, in consideration of the reasonableness of the test problems, the care taken in the eight years of test standardization, and the extent to which one agrees with the authors' premises and their criticisms of other intelligence tests.

Beyond its theoretical and research interest, however, a test must have some demonstrated relationship to some aspect of the school program to justify its widespread use for school purposes. Group intelligence tests now in use in schools have amply demonstrated effectiveness in predicting academic achievement, grouping pupils, aiding in vocational counseling, identifying underachievers, and the like. That they do these things imperfectly is obvious, but that they do them better than they could be done without tests is equally apparent.

In terms of practical considerations, then, if the Davis-Eells Test were to supplant presently used intelligence tests in the schools, it would need to be demonstrated that the test could perform the aforementioned functions better. If it were to supplement other tests, it should shed some new insights regarding specific children or groups of children. Since no single set of criteria appears available to assess its value in these roles, particularly the latter one, perhaps an accumulation of analytical studies may show ways in which the test could be used to advantage.

Problem

The intent of the present study was not to investigate the general validity of the Davis-Eells Games. Instead it poses several specific questions which, if the authors' premises be accepted, do not necessarily bear upon true test validity but are nevertheless relevant to the test's practical use in an elementary school setting: (*a*) What are the interrelationships among the Davis-Eells Test and the California Tests of Reading, Arithmetic, and Mental Maturity (3) for a typical sample of fourth-grade children? (*b*) Are the scores of bilingual children, which were not included in the standardization sample of the Davis-Eells Test though they are presumptively lowered by culture-bound items, significantly better on the Davis-Eells than on the California Test of Mental Maturity (CTMM)? (*c*) What relationships exist between social-class levels and scores on the above tests? (*d*) What types of children do better on the Davis-Eells Test than the CTMM, and vice versa?

Subjects

Since the Elementary A form of the Davis-Eells Games (hereafter referred to as the Games) is intended for use in grades 3–6, it was felt that choice of subjects from a fourth-grade sample would probably insure an adequate test ceiling as well as an adequate base for extreme scorers. Choice of fourth-graders was also a practical consideration since routine group achievement and intelligence testing at the fourth-grade level was already part of a county-wide testing program. Use of a relatively homogeneous age group also eliminated the need to partial out age as a factor in subsequent test correlations.

Children from all the fourth-grade classes of four elementary schools of Santa Barbara County participated in the study. The four schools, whose enrollments were between 300 and 700, were within a radius of ten miles from the city of Santa Barbara. Two of the schools drew their pupils from above-average residential areas, while the other two included a higher percentage of children of agricultural workers. The first two schools had few bilingual children, whereas in the latter two, approximately one child in five came from a home in which Spanish was spoken.

Prior test surveys in the four schools had suggested that, taken together, they would be a reasonable cross section of a fourth-grade school population. This later proved to be the case, with some minor exceptions to which reference will be made later.

A total of 184 children from the classes of nine teachers in the four schools constituted the final sample to whom all tests were given. There were 93 boys and 91 girls. These pupils represented 81 per cent of the total actual enrollment of all fourth-graders in the four schools. No children were intentionally omitted from the study, but absence on one or another of the testing days and other unavoidable circumstances prevented perfect participation.

Method of Test Administration

Four group tests were administered to all the children in their regular classrooms by their regular teachers. These tests were the Games, Elementary A; the California Test of Mental Maturity (CTMM), 1950 Elementary S-Form; the California Reading Test, Elementary AA; and the California Arithmetic Test, Elementary AA.

All tests were given during the first two months of the fall school term and rotated so that in some classes the Games were given first, in others the CTMM, and so on. Testing did not start until the term had been under way for several weeks. The two intelligence tests were given within a period of two weeks of each other in each class, and all tests in all four schools were administered within a calendar month and five days. Some time interval between tests was felt imperative by the teachers to avoid student fatigue and prevent undue interference with the school program.

Cumulative folders for all pupils were made available by the schools. Information relating to parental occupation, birthplaces of parents and child, language of the home, address, and date of child's enrollment in the given school were recorded when available.

Table 1

Characteristics of Sample with Respect to Age and Test Results

Item	Mean	*SD*
Age	9 yr. 4.3 mo.	6.4 mo.
Davis-Eells IPSA	101.0	15.5
Davis-Eells Raw Score	39.2	7.0
CTMM Total IQ	101.3	18.0
CTMM Language IQ	97.2	19.3
CTMM Non-Lang. IQ	107.4	22.7
Calif. Read. Gr. Pl.	4.1	1.4
Calif. Arith. Gr. Pl.	3.9	.8

Table 2

Pearsonian Correlations Between Measures of Intelligence

	CTMM IQ		
Test	Total	Lang.	Non-Lang.
Davis-Eells IPSA	.49	.37	.46
CTMM Total IQ		.86	.76
CTMM Language IQ			.42

Results

Test interrelationships for total sample. As far as mean scores on the two intellectual measures were concerned, the CTMM IQ and Games IPSA were virtually identical. Both were only a point above the norms of the standardization samples. Since the mean age of the children was exactly that of the California Test Bureau figure for fourth-graders in their second month of the school term, the sample may be considered representative in this respect also.

In several respects it deviates slightly from the norm, however. The standard deviation of IQ's for the CTMM and its subtests exceeds the norm, while that for the Games is less than the standard deviation indicated in several studies reported in the Manual (1).

More striking is the discrepancy between CTMM Language and Non-Language quotients favoring the latter. This is perhaps due in part to the scores of the bilingual children, whose relatively greater aptitude on nonverbal tests might be anticipated.

The reading-grade placements of the children, as can be seen from the table, were at the expected level. Their arithmetic scores were slightly below. The latter is perhaps due in part to differential rates of summer forgetting, with arithmetic scores the more likely to suffer from holiday disuse.[1]

Intercorrelations of the intelligence tests as shown in Table 2 indicate that the CTMM and Games are related less closely than most widely accepted intellectual measures. As one might have expected, the Games are more directly related to the CTMM Non-Language Test, which requires no reading, than to the Language Test, which does. Still the CTMM Total IQ has the highest *r* with the Games. The latter correlation of .49 is reasonably typical of those reported in the Manual for the Games vs. several forms of the Otis.

The relationships obtained from correlation of the Games with the achievement tests were also remarkably similar to comparable calculations reported in the Manual, where median coefficients of .43 for the Games vs. reading and .41 for the Games vs. arithmetic were reported for several batteries of achievement tests. It can be seen from Table 3 that the

Table 3

Pearsonian Correlations Between Measures of Aptitude and Achievement

Achievement Test Grade Placement	Davis-Eells Raw Score	CTMM Tot. MA	CTMM Lang. MA	CTMM N-L MA
California Reading	.48	.79	.77	.49
California Arithmetic	.43	.67	.58	.52

---

[1] It was originally planned that the reasoning and fundamentals subtests of the arithmetic test should be given separate statistical treatment and reported separately in all ensuing tables. It has been tentatively hypothesized that arithmetic fundamentals, not depending on reading as did the reasoning problems, might bear a greater degree of relationship to the Games and CTMM Non-Language than to the other tests. It had also been hypothesized that bilingual children might perform significantly better on the fundamentals test. Since neither of these hypotheses was supported by the data, only the total arithmetic score is reported herein.

CTMM Total has the highest degree of association with both indices of achievement. The CTMM Language has the next highest correlations, followed by the Non-Language which exceeds the Davis-Eells only slightly. Some changes in these correlations would have been effected by corrections for range, since the CTMM range was too wide and the Davis-Eells too narrow, reducing the correlations with the CTMM and increasing those with the Davis-Eells.

The bilingual sample. A bilingual child was herein defined as one whose school records included the written notation that a language other than English was spoken in the home. Such children, 31 in number, constituted 17 per cent of the total sample. Spanish was the second language of 27 of the children, Japanese of two, and German and Italian of one each.

It was interesting to find that all but three of the children were born in the United States and all but eight in Santa Barbara County. The majority of the fathers of these children had been born in the countries whose language they spoke, while the majority of the mothers were born in the United States. One might infer, then, that some English was also spoken around the homes of many of these children. Yet there was some presumptive language handicap operating which might be expected to influence adversely the score of the bilingual child on written tests.

Furthermore these children were also with few exceptions from underprivileged socioeconomic levels. Sixteen of their fathers worked as laborers, two as custodians, and the majority of the remainder in unskilled or ambiguously stated occupations. One barber, one painter, and one ranch foreman were the exceptions to the homogeneous parental occupational pattern of these bilingual children. The education of the fathers was also limited, and only five were listed as having gone beyond the junior high school level.

A comparison of the scores earned by the bilingual children on the various tests can be seen in Table 4. Their best intelligence test score was on the Games, followed closely by the CTMM Non-Language. As might be expected, the children's mean Language IQ was the lowest of the three. The difference of 11 IQ points between the Games IPSA and CTMM Language IQ is significant at the .01 level of confidence, but the 4-point difference between the Games and CTMM Non-Language had a critical ratio of only .74 and is, therefore, insignificant.

Table 4

Test Results for Bilingual Sample

Test	Mean	*SD*
Davis-Eells IPSA	97.7	15.4
CTMM Language IQ	86.7	14.3
CTMM Non-Language IQ	93.8	24.9
Calif. Reading Grade	3.7	1.3
Calif. Arith. Grade	3.7	.8

Parental occupation and measured intelligence of children. It had been hoped originally that it would be possible to obtain an Index of Status Characteristics for each child in the study, following the method described by Warner, Meeker, and Eells (4). Unfortunately, the information recorded in the available school records was inadequate for such a detailed appraisal. Even such items as parental occupation were frequently stated in an equivocal manner, e.g., "Self-employed." Nevertheless, parental occupation was decided upon as the best single available index of class level, and each child was given a rating of from 1 to 7 following the categories in the aforementioned reference.

The resultant distribution of scores was definitely overweighted in the professional and semiprofessional areas, with 17 per cent of the children so categorized. The parental occupations of 58 children unfortunately were either not given or not classifiable because of an inadequate description. Because of these marked limitations in the data, detailed statistical treatment seemed inadvisable. Nevertheless a rough grouping of the classes was made and mean IQ's computed. Results are shown in Table 5.

The fourth occupational class—skilled workers—had mean quotients of 105.0 and 103.9, respectively, on the CTMM and Games; including them in the grouping of the upper classes would have increased the significance of the difference between the test means to the .01 level of confidence. As it is, the dif-

Table 5

Parental Occupation and Intelligence Test Scores

Occupational class	*N*	CTMM IQ Mean	CTMM IQ *SD*	Davis-Eells IPSA Mean	Davis-Eells IPSA *SD*	Diff.	*CR*
Top 3 classes	41	113.1	10.2	106.4	13.4	6.7	2.55
Bottom 3 classes	67	95.6	16.1	97.9	14.0	2.3	.88

ference is significant at the .02 level, and in the expected direction: children from above-average backgrounds do better on the traditional type of intelligence test than on the Games. Yet the reverse does not hold true to a statistically significant degree; the children from the less privileged classes as here defined did not do significantly better on the Games than on the CTMM. The trends are in the expected direction but their magnitude is minimal.

The difference in variability in the two tests is also of interest. The difference between the CTMM standard deviations of scores for the top as opposed to bottom three classes is significant at the .01 level. That is, the top three classes have significantly less variability on the CTMM than do the bottom three classes. On the other hand, the variability of Games scores for the opposing occupational groupings is roughly the same. The CTMM scores appear closer to what one would expect of a "culture-bound" test.

The scores of children whose parental occupations were not given or were ambiguously stated had means of 98.4 and 100.1 on the CTMM and Games, respectively. It is possible that there was a disproportionate share of lower-class children in this group, but even including them with the lower groupings would have affected the findings but little.

Correlates of intelligence test discrepancies. Davis and Eells had indicated that problem-solving ability as measured by the Games was not to be strictly equated with standard concepts of intelligence as measured by most current group tests. It seemed of interest, therefore, to examine in some detail the characteristics of those children whose IPSA and IQ were markedly in disagreement, to see if meaningful correlates of the variable could be determined. A difference of 20 or more points between the scores obtained by a given child on the two tests was chosen rather arbitrarily as a cutting point which should indicate both statistical and diagnostic significance. Table 6 described the children who so excelled relatively on the Games and CTMM respectively.

One child in four had an IQ-IPSA discrepancy of 20 points or more. It can be seen that the children performing markedly better on the Games than on the CTMM constituted a slightly larger group than its opposite analogue: 14 per cent vs. 11 per cent of the total sample. As would have been hypothesized, more bilinguals excelled on the Games than CTMM.

The sex ratio is perhaps the reverse of what might have been anticipated in view of the stereotype of the greater verbalism of girls. More girls actually do better on the Games while more boys do better on the CTMM.

The test scores of the children excelling on

Table 6

Characteristics of Children with IPSA–CTMM Discrepancies of 20 or More Points

Item	IPSA Greater	CTMM Greater
Number of children	25	21
Number of bilinguals	9	1
Sex ratio, boys/girls	11/14	12/9
Mean Davis-Eells IPSA	112.8	96.0
Mean CTMM Total IQ	86.4	123.9
Mean CTMM Language IQ	82.0	122.0
Mean CTMM Non-Language IQ	92.0	124.0
Mean Reading-Grade Placement	3.6	5.3
Mean Arithmetic-Grade Placement	3.7	4.5
Parental occupation:		
Classes 1–3	2	4
Class 4	3	4
Classes 5–7	11	6
Unknown	9	7

the Games over the CTMM present, except for the IPSA, the test-score pattern of poor students. Their IQ's on the CTMM and its subportions are dull-normal to low-average; their achievement is below their grade placement in both reading and arithmetic. Yet in terms of the Games, the children are nearly a sigma above the mean in problem-solving ability.

Conversely, the 21 children with higher test scores on the CTMM are classified by the latter as superior in intelligence; they are also well above average in both reading and arithmetic. Had one evaluated these children by their mean problem-solving scores on the Games, however, they would have been described as slightly below average.

Both groups of children, nevertheless, have achievement scores that show a definite regression toward the mean from their CTMM predicted levels. By the latter, the slower group is overachieving and the brighter group underachieving. The former group is nearly a sigma below the mean in Total CTMM IQ, yet only about one-half sigma below in achievement. The latter group is nearly a sigma and a half above the mean in CTMM IQ, yet not a full sigma above in measured achievement.

Parental occupational levels follow to some extent the expected trend in that a higher proportion of upper-class children did better on the CTMM than on the Games, whereas a higher proportion of lower-class children did better on the Games than on the CTMM.

Summary and Conclusions

A group of 184 fourth-grade children, reasonably representative of fourth-graders in terms of age and measured intelligence, were given a battery of tests including the Davis-Eells Games, the California Test of Mental Maturity, and the California Tests of Reading and Arithmetic. Interrelationships among the tests were investigated, with special reference to the performance of bilingual children and those from varying social-class levels. While the general validity of the Davis-Eells Games cannot be inferred from the data obtained because of the test authors' insistence that academic performance is not an acceptable validation criterion, it may still be assumed from the relatively low correlations obtained that the Games should not be used in place of standard group intelligence tests for predicting school success, grouping pupils for instructional purposes, and similar traditional uses.

Comparison of the scores derived from the Games with those from the CTMM for bilingual children and children at various social-class levels suggested, though not conclusively, that children with a bilingual background and/or lower-class membership might do better on the Games than the CTMM. Trends were so slight that they need verification through other studies. Because of an atypical distribution of parental occupations, it was deemed inadvisable to compute correlations between socioeconomic status and aptitude. Havighurst (2) reports coefficients of .37 and .34 between the Games and socioeconomic status for fourth-grade girls and boys, respectively, from a midwestern community of approximately the same size as the one from which these data were obtained. The findings in combination suggest that whatever gains lower-class children make on the Games as opposed to other intelligence tests would be slight. Use of the test in elementary schools at present except as a research instrument does not appear to be warranted.

Received August 17, 1955.

References

1. Davis, A., & Eells, K. *Manual, Davis-Eells Test of General Intelligence or Problem-Solving Ability*. Yonkers, N. Y.: World Book Co., 1953.
2. Havighurst, R. J. (Ed.) Studying children and training counselors in a community program. *Suppl. Educ. Monogr.*, 1953, No. 78.
3. Sullivan, Elizabeth T., Clark, W. W., & Tiegs, E. W. *Manual, California Short-Form Test of Mental Maturity*. Los Angeles: California Test Bureau, 1950.
4. Warner, W. L., Meeker, Marchia, & Eells, K. *Social class in America*. Chicago: Science Research Associates, 1949.

Journal of Consulting Psychology
Vol. 20, No. 3, 1956

A Procedure for Obtaining Self-Ratings and Group Ratings[1]

Wilse B. Webb[2]
U. S. Naval School of Aviation Medicine

This is a report of the development of an effective technique for obtaining peer ratings and self-ratings of individual characteristics.

The procedure grew from research concerned with relations between group evaluations, self-evaluations, and objective measures (1, 3). In this research an odd obstacle developed. It was possible to obtain reliable group ratings by peer-nomination or rating-scale procedures for such characteristics as intelligence, leadership, personal charm, and the like. The difficulty lay in the self-ratings. Using either a self-ranking or a rating-scale procedure these ratings were quite unreliable. It was clear that until the reliabilities of the self-ratings could be increased the meaningfulness of our interrelations was limited.

As noted in a previous paper (3), manipulations of reliability formulas indicated that the high reliability of the group ratings accrued primarily from the number of ratings obtained on a given individual. Further, the lack of reliability of the self-ratings, using rating-scale or ranking procedures, yielded low reliabilities because of the "shortness" or "single item" characteristic of the rating.

There were two alternatives available for increasing the reliability of the self-ratings. One alternative is to turn to self-inventories in which a large number of *related* but different self-statements are made which are likely to be related to a given trait. For example, we could use the Guilford-Martin inventory to measure friendliness. On the other hand, we may attempt to obtain a large number of self-statements on the *same* question. For example, "How friendly are you?" could be asked a number of times. This procedure pursues the latter alternative.

For convenience's sake, the procedure is titled the "SPM Procedure." It will become clear in discussing the procedure that these initials stand for "self-plus-minus."

Procedure

To increase the number of statements that the man made about himself, a method of paired comparison was introduced. Very simply, the individual was given a roster of the individuals within his group. He was asked to go through this roster and compare himself with every other man in the group. If he considered himself superior on a particular trait to a given man, he assigned a plus by that name. If in contrast he considered himself inferior on a given trait, he assigned a minus to that name. His self-rating rank within the group then could be given by simply counting the number of pluses which he assigned to the members of his group and subtracting from the total N. If, for example, he had given 15 pluses in a group of 24 men, he considered himself to be superior to 15 men in that group and 8 men to be inferior to him, and his rank on that particular trait from his point of view was 9 in the group $(24-15 = 9)$.

In working with this SPM procedure, an additional benefit was noted. By summing the total number of pluses assigned an individual by all members of the group (the number of men who consider themselves superior to an

---

[1] Opinions and conclusions contained in this report are those of the author. They are not to be construed as necessarily reflecting the view or the endorsement of the Navy Department.

[2] My thanks are due to and freely given to Edward J. Wallon for his help in data collecting and calculations.

Table 1

A Model of the SPM Rating Procedure for Five Cases

	A	B	C	D	E	Group ranks
A		−	−	−	+	2
B	+		−	+	+	4
C	+	+		+	+	5
D	+	−	−		+	3
E	+	+	−	+		4
Self-ranks	1	3	5	2	1	

+ = column designee considers himself superior to row designee.

− = column designee considers himself inferior to row designee.

Self-ranks = 5 − pluses of each column.

Group ranks = 1 + pluses of each column.

individual) and by adding 1 a ranking of the group judgment of this man could be obtained. An example is given in Table 1. Let us assume that there are five men in a given group. The columns of Table 1 represent the pluses and minuses assigned by each individual to the other members of the group on a given trait. The score at the end of each column (5-the sum of the pluses) indicates the man's self-rank which he assigns himself. When the table is constructed in this manner, it is quite apparent that the rows can represent the judgment of a given individual by his entire group and that summing across the rows the relative ranks of individuals as judged by his fellow members in the group on a given trait may be obtained. If, for example, as in the case of man *A*, 3 members of the group considered him superior to themselves, he is clearly a superior man in terms of group judgments on this trait. If in contrast (as in the case of man *C*) all members considered themselves superior to him, he is then considered an inferior member in this particular group. In the case of man *E*, he considers himself superior to all members of the group whereas only one member of the group rates him as superior.

Results

Eight sections have been tested using the SPM procedure in the fourth week and fifteenth week of preflight training of the Naval Air Training Program. These sections ranged in number from 21 to 34 members. The characteristics of intelligence and leadership were used. Estimates of the reliability within a given session and between sessions for the two characteristics were obtained for both the group ratings and the self-ratings. These data are summarized in Table 2.

The between-session reliabilities of the ratings given in Table 2 are the averages of the product-moment correlations for each of the eight sections between the fourth- and fifteenth-week administration of the forms. The within-session reliabilities are the averaged reliability estimates within the fourth- and fifteenth-week administrations obtained for each section. These estimates were made by an analysis-of-variance procedure (comparing the row and column variances with the remaining error variance). The average number for each section in the fourth week was 28.7 and for the fifteenth week the average *N* was 25.9.

These reliability estimates may be compared with the reliabilities of group ratings obtained by peer-nomination procedures and rating-scale procedures from data on comparable populations. The widely used peer-nomination technique has been applied to a number of cadet populations. The group members were asked to write the names of the three "highest" in their section on leadership potential and also the three "lowest" on leadership potential. Group measures were developed by assigning a plus when a man was nominated "high" and a minus when a man was nominated "low" and algebraically summing the pluses and minuses for each individual within the group. An average within-session reliability of this technique for six sections of cadets has been reported as .88 (2). The total number of cases in the study

Table 2

Reliabilities Within and Between Sessions for SPM Groups and Self-Rating on Intelligence and Leadership

Week	Intelligence group rating	Intelligence self-rating	Leadership group rating	Leadership self-rating
4th	.90	.91	.92	.90
15th	.92	.90	.92	.88
4th–15th	.76	.58	.75	.69

was 116. A between-fourth-and-fifteenth-week session reliability on three sections of cadets gave a reliability estimate of .78 for a total of 68 cases (unpublished data). In both of these instances the trait of leadership was considered. Group ratings have also been obtained by a 7-point rating-scale procedure in which each man rated every other man in his section and himself on intelligence. This procedure yielded within-session reliabilities of .86 and .89 at the fourth week and fifteenth week, respectively, and a between-session reliability of .76 (3).

The reliability of two methods of obtaining self-ratings may also be compared with the SPM procedure. A between-session reliability of leadership rankings (each subject was asked to assign himself a rank number within the total number of people in his section) was obtained for three sections of cadets. The average between the fourth- and fifteenth-week administrations for the three sections of cadets (total *N* of 68 cases) was .34 (unpublished data). The rating-scale procedure (each individual rating himself on a 7-point scale) yielded a between-session of .19 for 95 cases (3).

The reliability findings may be summarized. The group-rating reliabilities both within and between sessions for three methods are quite comparable. The SPM procedure enjoys a considerable advantage in the self-rating reliabilities between sessions.

The comparability of the SPM procedure for group ratings and the peer-nomination procedure was determined. The two alternative procedures were used in the fifteenth-week sections in which SPM data of Table 2 were obtained. The averaged correlation between the group ratings obtained by the two procedures was .80. In addition to the peer-nominations the rank-order procedure for obtaining self-ratings was also introduced at the fifteenth week for the eight sections. These rankings correlated .34 with the SPM self-rating ranks.

The relationships between the measures obtained by the SPM procedure with objective measures can be roughly compared with those relationships obtained by the rating procedure on the variable of intelligence. This comparison, however, is only approximate in that the object measure for the SPM method is the ACE and in the case of the rating-scale study, this measure was the Otis self-administering test (3). The data are given in Table 3.

Table 3

Correlations Between SPM Method and the ACE and Rating-Scale Method and Otis for Group Rating and Self-Rating

	Group rating	Self-rating
SPM	.46	.26
Rating Scale	.49	.21

Summary

Summary statements of the available statistical data concerning the SPM procedure and its comparison to other methods can be given:

1. The SPM procedure yields highly reliable measures within sessions for both group ratings and self-ratings.
2. The between-session reliabilities of the group ratings are high and the between-session self-ratings are respectable.
3. The SPM procedure increases the reliability of between-session self-ratings when compared with self-ratings given by rank designation procedure or a self-rating on a rating scale.
4. The group ratings obtained by the SPM method are considerably correlated with the peer-nomination procedure for group ratings.
5. The self-ratings of the SPM procedure within sessions show only a limited relation with a ranking procedure. Comparing the between-session reliability would favor the use of the SPM method.
6. The group ratings show substantial and similar relations to objective measures when compared with those obtained by a rating procedure.
7. Self-ratings show low and similar relations to objective measures when compared with those obtained by a rating procedure.

In addition to these statistical facts, a number of "fringe benefits" may be cited for the procedure outlined in this paper:

1. Measures of intrasession reliability are available for the self-ratings.

2. Group ratings are obtained indirectly without the man being forced to name high or low extremes of his group or to rate individuals extremely.

3. The procedure is extremely simple to administer and requires a shorter time than the other procedures. It can be easily administered on mark-sensing cards or on IBM answer sheets. We use the later procedure in which each name of the class roster is assigned an answer space and the rater marks answer space "a" if he considers himself superior for a given item (name–number) or "b" if he considers himself inferior.

4. There is a common set for all subjects in the case of the self-ratings. For example, the other procedures permit the individual to rate as he thinks the others will rate him or as he thinks he should be rated.

5. The value of a trait to a group may be obtained indirectly by counting the total number of pluses assigned a trait by the group. Indirect measures of morale may be obtained through such a plus-minus counting for the total group.

Received August 24, 1955.

References

1. Webb, W. B. Self-evaluation compared with group evaluations. *J. consult. Psychol.,* 1952, **16,** 305–307.
2. Webb, W. B. The reliability of peer ratings. *U. S. N. Sch. Aviat. Med. Res. Rep.,* No. NM 001 058.25. 13. 20 April 1954.
3. Webb, W. B. Self-evaluations, group evaluations, and objective measures. *J. consult. Psychol.,* 1955, **19,** 210–212.

PSYCHOLOGICAL TEST REVIEWS

New Tests

Bennett, George K., & Gelink, Marjorie. *Short Employment Tests (SET), 1956 Manual.* New York: Psychological Corp., 1956. Pp. 11.

The second revision of the manual for the *Short Employment Tests* (see *J. consult. Psychol.,* 1952, **16**, 159; 1953, **17**, 401) contains correlations of the tests with job criteria in 19 business situations. The norms for job applicants are now based on more than 33,000 women and 9,000 men.—*L. F. S.*

Brown, William F., & Holtzman, Wayne H. *Brown-Holtzman Survey of Study Habits and Attitudes (SSHA), 1956 Manual.* New York: Psychological Corp., 1956. Pp. 11.

The revised manual contains technical data for high school samples as well as for college students (see *J. consult. Psychol.,* 1954, **18**, 153–154). At the high school as well as at the college level, the SSHA shows substantial correlation with achievement, has low correlation with the ACE, and adds significantly to the multiple correlation. Percentile norms are given for college and high school groups, derived from 3,500 and 2,800 cases, respectively.—*L. F. S.*

Buhler, Charlotte, & Manson, Morse P. *The Picture World Test.* Individual projective technique. Child-adolescent-adult. 1 form. Untimed, (30) min. Set of materials ($18.50 for 25 administrations), with manual, pp. 86; manual only ($3.50). West Los Angeles, Calif.: Western Psychological Services, 1956.

The *Picture World Test* is essentially a picture version of the well-known *World Test,* designed to explore how a person perceives and structures his world, with emphasis on his goal-setting activities. The examinee is presented with 12 small drawings of structured scenes, a large sheet of blank paper, and a chart of 36 simple symbols by which he can add persons, vehicles, buildings, and animals to the scenes. He is asked to choose any number of the scenes he wishes "to make up a world as it is or as you would like it to be; the world you like or dislike; the world of your dreams. . . ." The examinee gums the scenes to the blank sheet, connects the scenes as he wishes, adds symbols, gives his world a name or title, and tells a story about it.

The manual describes methods for interpreting the world scenes and world stories. Some data are given on the frequencies of the main story categories in relation to age and quality of adjustment. Twenty-eight illustrative case studies demonstrate the clinical use and interpretation of the test.

All in all, the *Picture World Test* seems to be an interesting and challenging instrument, likely to evoke self-revealing responses of considerable clinical value.—*L. F. S.*

Cureton, Edward E., Cureton, Louise W., *et al. The Multi-Aptitude Test.* For demonstration in classes and lay groups. 2 forms. 35 (50) min. Study kit, with 2 test forms, keys, and manual, pp. 32 ($1.25 ea., $9.00 per 10); test booklet separately, form A or B ($3.50 per 25). New York: Psychological Corp., 1955.

The Cureton *Multi-Aptitude Test* is unique. Not intended for any practical purpose of measurement, it is designed to teach and demonstrate the nature of psychological tests to a wide range of interested persons. Advanced classes may use the test data for item analyses and studies of test intercorrelations; nonprofessional groups may satisfy their curiosity about what tests are. The 10 short subtests represent most of the common varieties of test item. Five parts—vocabulary, information, arithmetic, number series, and figure classification—are typical of group mental tests and also illustrate verbal, number, and reasoning factors. The remaining subtests—mechanical comprehension, word recognition, scrambled letters, checking, and paper form board—typify other abilities found by factorial studies and used in special aptitude tests. The keys acquaint students with various types of scoring devices. The manual is an excellent miniature specimen for instruction, with full tables of data and norms for practice, although not for practical interpretation. The test should be widely useful as an unrestricted instrument for instruction, thereby sparing other tests whose security needs protection.—*L. F. S.*

Phillipson, Herbert. *The Object Relations Technique.* Individual projective technique. 1 form. Adolescent-adult. Average time (90) min. Set of 13 plates, with manual, pp. x + 224, cloth ($10.00); manual only ($6.00). Glencoe, Ill.: The Free Press, 1955.

The *Object Relations Test* is a new projective method which bears some resemblance to the TAT, developed at the Tavistock Clinic at London. Its novel features are its rationale which springs from a theory of personality, and its stimulus pictures which were designed so as to vary certain important characteristics systematically. The test's name and intent come from the psychoanalytic theory of unconscious object relations which exist within the personality, were derived from the individual's relations with significant persons in his earlier life, and in turn determine his relationships with the real people in his present external world.

Twelve of the pictures—the thirteenth is a blank card—are divided into three series of four plates each. The series vary in degree of structure, from vague, hazy figures and backgrounds in series *A*, to detailed, realistic backgrounds in series *C*. In all series, the human figures are ambiguous as to sex and lack facial features. The four cards within each series represent one figure, two figures, three figures, and a group. Administration is similar to that of the TAT. The cards are not shown in serial order, but in a designated order which mixes the series and the numbers.

The accompanying book discusses the theory on which the test is based, describes the materials and their administration, and gives methods for the analysis and interpretation of the stories. The interpretation of one protocol is given at length, and six other cases are presented more briefly. Normative information is given for 50 clinic outpatients and for a sample of 40 normal adolescent girls. Although such data may seem scanty if judged by standards appropriate to mass testing, they exceed the material usually offered with a new projective method and are unquestionably useful.

The pictures are sensitively conceived, and their systematic plan commends them both for clinical use and for research. This instrument deserves a thorough exploration by American psychologists and may well prove to be a major development in projective methods.—*L. F. S.*

Segel, David, & Raskin, Evelyn. *Multiple Aptitude Tests*. Grades 7–13. 1 form. 175 (210) min. 9 test booklets ($2.45 to $4.55 per 35, each; set, $24.50 per 35), with manual, pp. 96, keys; 2 IBM answer sheets (4¢); extended or transparent profile sheets (2¢); sample set ($1.75). Los Angeles: California Test Bureau, 1955.

The *Multiple Aptitude Test* offers another battery of differentiated tests for high school students and college freshmen. The nine subtests—word meaning, paragraph meaning, language usage, clerical facility, arithmetic reasoning, arithmetic computation, applied science and mechanics, spatial relations in two dimensions, and spatial relations in three dimensions—may be combined to yield four factors, verbal, perceptual speed, numerical, and spatial. The 96-page manual gives unusually full data, although there are some conspicuous gaps. Reliabilities by Kuder-Richardson formula 21 range, in single-grade groups, from .72 to .92 for the subtests and from .91 to .95 for the factors. Validity is discussed in terms of the factor analysis, correlations with other tests, and correlations with school marks in 18 subjects. The norms, obtained on a nationally distributed sample of over 11,000 cases in grades 7 through 13, are expressed as standard scores and percentile ranks for each subtest.

The data in the manual reveal a number of interesting strengths. Good attention is paid to the standard errors of differences between scores, and users are cautioned against drawing conclusions from nonsignificant differences. An examinee's test profile can be compared with those of high and low achievers in each school subject, and expectancy tables are provided which relate school marks to scores on the most predictive subtests.

There are also some weaknesses. Most striking is the fact that all norms and interpretive data are given in terms of the nine subtest scores, while the more reliable combined factor scores are almost ignored. The manual gives little information about the selection of content and the development of the subtests. The data also disclose some perplexities. Why is the arithmetic computation subtest not only the best predictor of grades for general mathematics and general science, but also the best or close next-best for English, foreign languages, social studies, and psychology? Such an observation casts some doubt on the value of the tests for differential prediction in comparison to a single predictive score.—*L. F. S.*

Wagner, Mazie E., & Schubert, Herman J. P. *D. A. P. Quality Scale for Late Adolescents and Young Adults*. Buffalo 22, N. Y. (State University of New York College for Teachers): Authors, 1955. Pp. 23 + 28 plates. $2.00.

A long-established way to evaluate a personal product globally is by the use of scaled specimens for comparison. Wagner and Schubert have produced four scales for rating drawings of a person—male and female figures, drawn as front and side views. Data in the accompanying monograph show that trained raters can use the scales with high reliability (> .90), that novices can learn to rate quickly, and that examinee retest reliability for drawings so rated is .85. Most interesting is the evidence that the drawing scores contribute significantly to the prediction of scholarship in a teachers' college, and have low correlations with other predictors. The findings suggest that the DAP may be used to predict the achievement of young adults, just as Goodenough used it to predict the achievement of young children many years ago. The scales also demonstrate the rewards which come from applying sound quantitative methods to the over-all appraisal of a complex personal product.—*L. F. S.*

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

Validation of the Lowenfeld Mosaic Test[1]

Monroe L. Levin
Teachers College, Columbia University[2]

The study upon which this report is based was stimulated by personal experience with the Mosaic Test which had led to the impression that it was a useful clinical instrument. Although the test has been applied to many clinical problems and has been endorsed by numerous articles, a critical review revealed the literature to be contradictory, diffuse, and inconsistent. Furthermore, such basic issues as scoring reliability and experimental validation of hypotheses had never been adequately investigated.

Method

The double set of the Lowenfeld Mosaic Test was administered to all subjects according to the standardized instructions recommended for use in this country (18). All subjects were limited to a performance time of twenty minutes. The final Mosaics were photographed on 35 mm. Type A Kodachrome Film yielding color transparencies which were readily examined by means of a slide projector.

Hypotheses

Hypotheses were in the form of statements of expectancy which stated that a given Mosaic sign would occur significantly more often in the experimental group being examined. Hypotheses were formulated which were expected to differentiate between: (*a*) institutional and noninstitutional groups, (*b*) between normals and all other subjects, (*c*) between "maladjusted" and normal subjects, (*d*) among six diagnostic groups, and (*e*) between subjects with high scores in the clusters obtained from the Wittenborn Psychiatric Rating Scales and those with low scores in the same clusters. Hypotheses were also tested which had been suggested or inferred by Lowenfeld and Wertham but which had not been originally included in the study. Most of the hypotheses were based on the published Mosaic literature.

Institutional and noninstitutional groups. The Mosaic Test has been regarded as a Gestalt instrument which reflects the functioning level of the individual's cognitive and emotional processes in concrete situations (16, 21). Since the incoherent Mosaic is by definition the most disorganized response to the test, it was anticipated on an a priori basis that incoherent designs would occur much more frequently in institutional groups than in noninstitutional groups which are able to function in society, regardless of diagnostic considerations. It was therefore hypothesized that when the institutional groups were compared with the noninstitutional groups, incoherent designs would be found to occur significantly more often in the former group.

Success and adjustment. It was anticipated that when normal and pathological groups were compared, unsuccessful designs would occur more frequently among the pathological groups since lack of success has generally been regarded as a gross sign of psychopathology (2, 6, 10, 12, 16, 17). It was also anticipated that the proportions of occurrence would be such that the sign could be used in making individual discriminations.

"Maladjusted" vs. normal subjects. Subjects who were functioning in society were divided into "normal" and "maladjusted" groups on the basis of scores on the Cornell Index (Form N2). Since the "maladjusted" group was heterogeneous, it was impossible to develop hypotheses which could have been expected to identify a majority. It was possible, however, to select Mosaic characteristics which had been regarded as being indicative of the presence of

---

[1] Based upon a doctoral dissertation submitted to Columbia University in 1954. The author is indebted to professors L. F. Shaffer, E. J. Shoben, Jr., and H. Solomon for their assistance and guidance. The research was performed while the author was a Research Fellow of the National Institute of Mental Health, United States Public Health Service.

[2] Currently at Western State Hospital and Christian County Mental Health Center, Hopkinsville, Kentucky.

a wide variety of emotional disturbances such as might be present among the "maladjusted" group.

Twenty characteristics were selected. It was reasoned that if the twenty signs have utility in discriminating between better and less well adjusted socially functioning individuals, they would occur significantly more often among the "maladjusted" group. The signs which were selected as hypotheses and the sources for each are given in Table 2.

After the data were analyzed, it was decided to determine whether, as anticipated, "maladjusted" subjects produced more of the individual signs than did normals: i.e., whether the sum of the signs was discriminative.

Diagnostic groups. In formulating hypotheses for the diagnostic groups, signs were combined to form a multiple hypothesis when they could be expected to occur in about the same proportion of the sample. Experience with the Mosaic Test was relied upon to construct final hypotheses which seemed most likely to yield positive results. The specific hypotheses and their sources are given in Table 3. The hypothesis that psychotic subjects would use fewer pieces than nonpsychotic subjects was also tested.

Wittenborn cluster groups. Since the Mosaic literature rarely differentiated between specific types of neurosis or psychosis, it was necessary to devise a procedure for assigning hypotheses to the Wittenborn clusters. It was possible to assign some Mosaic signs of neurosis and psychosis to specific "neurotic" and "psychotic" cluster groups. Additional hypotheses were formulated by analyzing the Wittenborn Scales in terms of the Mosaic characteristics which the experimenter believed could be expected from individuals representing the extreme of behavior on which each of the scales is based. It was then possible to obtain "Mosaic clusters" which were hypothetically directly related to the Wittenborn clusters. Table 4 lists each of the specific hypotheses and their sources. It was also hypothesized that subjects at the upper end (most pathology) of the depressed cluster group would have a slower reaction time than other groups.

Lowenfeld's and Wertham's signs. Those signs of Lowenfeld and Wertham which had not been included as hypotheses or else had been modified drastically, but which were examined after the data had been gathered, are indicated in Tables 5 and 6, respectively.

Subjects

Since most of the Mosaic literature is concerned with descriptions of the performances of diagnostic groups, and statements about the usefulness of test characteristics for differential diagnosis, psychiatric diagnosis was the first criterion with which Mosaic Test performance was compared. In addition to the groups which were defined as "normal" and "maladjusted," four institutionalized groups were included in the study. The institutionalized groups consisted of diagnosed neurotics, familial mental defectives, general paretics with psychoses, and schizophrenics.

Because of the unreliability of psychiatric diagnosis, the Wittenborn Psychiatric Rating Scales were used to select groups for the second phase of the research.

Selective criteria. All subjects were white males between the ages of twenty-one and fifty-nine. Fourteen of the normals, three of the "maladjusted," four of the neurotic, seven of the paretic, and one of the schizophrenics were foreign born.

All of the noninstitutional subjects were functioning in society at the time of testing. They were drawn from a variety of residential and occupational settings in New York City. The group which was defined as normal consisted of those noninstitutionalized persons whose responses to the Cornell Index (Form N2) resulted in scores of twelve or less with no "stop" items, while the group defined as "maladjusted" consisted of individuals whose scores on the Cornell Index exceeded the assigned cutoff level. Persons whose indices revealed that they had histories of mental illness, ulcers, convulsions, asthma, drug or alcohol addiction, were dropped from the study, and those known to be in psychotherapy at the time of testing were also excluded.

None of the institutionalized subjects had ever had a previous diagnosis which conflicted with the current one: e.g., none of the neurotics or defectives had been previously diagnosed as being psychotic. With the exception of some of the neurotics, there was no staff disagreement concerning the adequacy of the diagnoses. It was necessary to disregard staff differences and use the official diagnoses in order to obtain a sample of diagnosed neurotics. It is possible that some of the neurotic subjects were in the initial stages of psychoses.

Subjects in the general paretic group had had no history of mental illness prior to the onset of leutic symptoms, and none of the familial mental defectives manifested noteworthy symptoms of emotional disturbance. There were 52 normal, 14 "maladjusted," 14 neurotic, 34 defective, 29 paretic, and 35 schizophrenic subjects.

It was possible to obtain ratings of 35 schizophrenic, 29 general paretic, and 10 neurotic subjects on the Wittenborn Psychiatric Rating Scales. Groups were selected to represent the upper and lower extremes of each Wittenborn symptom cluster by means of scores on each of the nine clusters. Subjects at the lower end (least pathology) of each cluster were compared with subjects at the upper end (most pathology) of that cluster. Because of sampling difficulties, two sets of samples were selected from the upper ends of some of the Wittenborn clusters: a smaller sample in which the extremes of pathology were most marked, and a larger sample in which the extremes of pathology were somewhat attenuated.

Uncontrolled variables. It was not possible to control the variables of socioeconomic status, intelligence, age, and ethnic origin. Attempts were made to avoid overrepresentation of any one socioeconomic group, although if Mosaic Test responses are primarily determined by personality characteristics, socioeconomic status should not be a major determiner of test responses.

There has been a consensus that, after the Mental Age of eight years, age is a relatively unimportant factor in Mosaic performance (2, 14). Conversely, it might be stated that after MA 8–0, intelligence is a relatively unimportant factor in Mosaic performance. It was decided to allow intelligence to remain uncontrolled because of the difficulty of obtaining accurate measures from certain of the pathological samples. It seemed likely that intelligence would be a minor variable when compared with personality, particularly since sampling methods probably resulted in a wide range of intelligence levels among every group except the mentally defective one.

Two auxiliary studies demonstrated that lack of control of the variables of age and ethnic origin did not seriously affect the results (8).

Scoring Method

In a preliminary experiment, a method for scoring Mosaic characteristics was devised. Scoring of 29 Mosaics by three independently trained judges and the author yielded proportions of agreement on Mosaic characteristics which ranged from .45 to 1.00. Chance alone would have yielded a ratio of .125 perfect agreements in any instance. Additionally, tests of concordance revealed that the judges never deviated from an arbitrary standard which required them to agree at least 90 per cent of the time. It was therefore concluded that the method for scoring Mosaic characteristics was teachable and resulted in high interjudge agreement.

Steps were taken to eliminate bias in scoring of the Mosaics in the research proper. To determine whether bias had been successfully controlled, and to re-examine scoring reliability on the research data proper, a layman was trained to score Mosaics. She then scored the 20 Mosaics considered to have been most difficult to score. Percentages of agreement between the layman judge and the author ranged from 75 to 100 per cent on the hypothesized signs. In every instance the obtained agreements differed significantly from chance expectancy. It was concluded that bias in scoring had been successfully controlled and that the reliability of the devised method of Mosaic scoring was independently reconfirmed (8).

Reliabilities of combinations of signs. Most of the hypotheses consisted of several signs used in combination, while the reliability data were based almost entirely on single Mosaic signs. Although the multiple hypotheses should have lower scoring reliabilities than the individual signs, the scoring reliabilities which were obtained were so high that it was assumed that the combined signs would have satisfactory scoring reliabilities.

Statistical Methods

Frequencies of occurrence of hypothesized signs were often so small that minor changes in frequencies could have seriously modified the statistical conclusions. The 1 per cent level was set as the criterion of statistical significance to offset the effects of chance sampling fluctuations.

Descriptive method. Values of $\sqrt{\chi^2/N}$ were computed in those cases where subjects were not selected independently of each other. The method was appropriate for examination of results obtained when Wittenborn cluster

Table 1

Notation Used in Descriptive Analysis of Data

	Wittenborn Cluster Score	
Sign	Low	High
Mosaic sign absent	A	B
Mosaic sign present	C	D

Table 2

Frequencies of Occurrence and Values of z for Twenty Signs of Maladjustment When Maladjusted and Normal Groups are Compared

Item	References	Maladjusted f	Normal f	z
1. Edge	3, 7, 9, 10, 14, 16, 17, 19	3	12	−1.17
2. Corner	7, 9, 10	0	1	− .92
3. Frame	1, 7, 9, 10, 14, 16, 17, 19	2	3	.96
4. All-over	2, 3, 16, 17, 19	1	6	− .51
5. Well-defined cross	9	0	3	−1.62
6. Supersymmetry	2, 6, 16, 17	0	0	0
7. Active C rejection plus emphasis on F	2, 16, 17	6	11	1.57
8. Emph. on C without emph. on F	2, 16, 17	0	2	−1.32
9. Unsuccessful	2, 6, 10, 16, 17	9	14	2.56
10. 40% or more black in comb. with other Cs. No others preponderant	9, 14, 16, 17	1	0	1.79
11. 40% or more black and blue in comb. with other Cs. No others preponderant	16, 17	5	13	.78
12. 40% or more red	16, 17	0	6	−2.3
13. Serrated red edge	9	1	0	−1.79
14. Superimposition of stones	2, 14	0	4	−1.89
15. Attempted edging	2	0	1	− .92
16. Winged designs	7, 9, 16, 17	0	3	−1.62
17. Downward arrow	19	1	5	− .3
18. Repetitive simple design	2, 16	0	0	0
19. Formation of letters of alphabet	—	0	3	−1.62
20. Use of full 20′ with resultant unsuccessful design	—	0	2	−1.32

groups were examined. The notation used is indicated in Table 1.

$$\sqrt{\chi^2/N} = \sqrt{(A - D)^2/(A + D)N}$$

was the statistic which was used. The closer the value of $\sqrt{\chi^2/N}$ to 1.00, the better the agreement between the Wittenborn Scales and the Mosaic Test in selecting individuals who showed and failed to show the clusters of behavior which were being considered.

Results

Analyses of the data revealed that only one hypothesis functioned as anticipated, although some signs which had been expected to discriminate between one group and others did so in a few, rather than in all instances.

Institutional and Noninstitutional Groups

When the 66 socially functioning subjects were compared with the 112 institutionalized subjects it was found that incoherent Mosaics were produced by four of the former and twelve of the latter subjects. Arc sine analysis of the data yielded a z of -1.08, indicating that the data clearly failed to support the hypothesis being tested.

Success and Adjustment

In scoring for "success" both the intent of the subject and his final Mosaic production were scored. In all other instances in this study, only the final product was considered. Fourteen of the 52 normal subjects and 89 of the 126 less well adjusted and endowed subjects produced unsuccessful designs, resulting in a z of 5.67. Although the statistical data strongly supported the hypothesis that unsuccessful designs occur most frequently among poorly adjusted and poorly endowed subjects, the frequencies of occurrence were such that the sign would not be useful in making individual discriminations. Since success and lack of success have perfect negative correlation by definition, the statistical results also supported the implicit hypothesis that successful designs occur most frequently among well endowed, well adjusted subjects.

"Maladjusted" vs. Normal Subjects

The results obtained by comparing the performances of normal and "maladjusted" subjects are summarized in Table 2. None of the signs of maladjustment discriminated between normal and "maladjusted" subjects at the 1 per cent level of significance. A subsequent analysis of the data revealed that normal subjects produced a mean of 1.73 signs of maladjustment, and "maladjusted" subjects produced a mean of 2.07 such signs. Since the standard error of the difference between means was .3 and *t* had a resultant value of 1.12, it was concluded that the number of signs produced by normal and "maladjusted" subjects did not differ significantly. The latter finding is in need of cross validation, since the hypothesis was not formulated in advance.

Diagnostic Groups

Examination of Table 3 reveals that none of the hypotheses relating to the diagnostic

Table 3

Frequencies of Occurrence, Tests of Significance, and References for Hypothesized Signs Related to Diagnostic Groups

			Exp. Group	Comparison Groups				
Sign	References		Neurotic	Defect.	Paretic	Schiz.	Maladj.	Normal
Corner	9, 10	*f*	0	0	1	0	0	1
		z	—	0	−1.15	0	0	− .92
Frame	9, 10, 16	*f*	1	1	0	1	2	3
		z	—	.62	1.65	.63	.62	.18
Cross	9	*f*	0	1	1	2	0	3
		z	—	−1.08	−1.15	−1.53	0	−1.62
40% or more black in comb. with other Cs. No others preponderant	9, 14, 17	*f*	0	1	0	3	1	0
		z	—	−1.08	0	−1.89	−1.42	0
Serrated red edge	9, 16	*f*	0	0	2	1	1	0
		z	—	0	−1.63	−1.49	−1.42	0
Winged design	7, 9, 19	*f*	1	0	0	0	0	3
		z	—	1.7	1.65	1.71	1.42	.18
			Defect.	Neurotic	Paretic	Schiz.	Maladj.	Normal
Unembellished fundamentals	9, 19	*f*	1	0	1	0	0	1
		z	—	1.08	−.13	1.42	1.08	2.98*
Simple design	9, 13, 15, 19	*f*	31	1	24	23	7	15
		z	—	6.3*	1	1.04	3.05*	6.36*
No. small, simple compact designs	9, 13, 15, 16, 17	*f*	3	0	0	0	0	0
		z	—	1.9	2.38	2.5	1.9	2.73*
Repetit. design of paired stones	9, 19	*f*	3	0	0	0	0	0
		z	—	1.9	2.38	2.5	1.9	2.73*
Enumerative use of color	16, 17	*f*	7	1	5	4	1	0
		z	—	1.27	.34	1.05	1.27	4.26*

* Difference significant at 1% level.

Table 3—*Continued*

Sign	References		Exp. Group	Comparison Groups				
			Paretic	Neurotic	Defect.	Schiz.	Maladj.	Normal
Unsuccessful, simple des., C used indiscriminately	2, 12, 16, 17	*f*	11	0	12	8	6	6
		z	—	4.18*	.21	1.36	−.32	2.87*
Inappropriate use of shape	12, 16, 17	*f*	4	0	3	0	1	2
		z	—	2.33	.63	3.03*	.11	1.56
			Schiz.	Neurotic	Defect.	Paretic	Maladj.	Normal
Emph. on F with active C rejection	2, 6, 16, 17	*f*	12	3	13	16	6	11
		z	—	.92	−.34	1.22	−.56	1.34
Supersymmetry	2, 16, 17	*f*	6	1	1	1	0	0
		z	—	1	2.12	1.9	2.71*	3.88*
W. tray covered by complex incoherent des. begun within 10″	2	*f*	0	0	0	0	0	0
		z	—	0	0	0	0	0
Edging	2	*f*	0	0	0	0	0	1
		z	—	0	0	0	0	−.88
Superimposed stones	2	*f*	6	0	1	4	0	4
		z	—	2.71*	2.12	.36	2.71*	1.32
Repetit. simple des. with same shape repeated, variations in C	16	*f*	0	0	0	1	0	0
		z	—	0	0	−1.5	0	0
Abstract des. predominate	16, 17	*f*	3	5	9	10	3	9
		z	—	−2.18	−2.02	−1.94	−1.17	−1.2

* Difference significant at 1% level.

groups functioned as predicted. The data failed to support the additional hypothesis that psychotic subjects would use fewer pieces than nonpsychotic subjects. Psychotic subjects used a mean of 33.3 pieces with a standard deviation of 43.87; all nonpsychotic subjects used a mean of 28.45 pieces with a standard deviation of 23.39; and normal subjects used a mean of 32.64 pieces with a standard deviation of 40.36. It is therefore obvious that there were no significant differences between the numbers of pieces used by psychotic, nonpsychotic, and normal subjects.

Wittenborn Cluster Groups

Table 4 summarizes the data relative to the Wittenborn cluster groups. Although examinations of Columns 1 and 2 demonstrate that the Mosaic Test usually correctly identified individuals who did not demonstrate the symptom correlates of the clusters, comparisons of Column 3 and 4 show that the test rarely identified individuals who demonstrated the extreme behaviors associated with the Wittenborn clusters. Values of $\sqrt{\chi^2/N}$ computed from the data in Columns 1 and 4 indicate that the Mosaic Test and Wittenborn Scales were generally in good agreement in identifying pathological and nonpathological individuals, but this measure did not consider instances (Columns 2 and 3) where the Mosaic Test failed to function as anticipated. Were it possible to include all the data in the statistical computations, it is probable that tests of the null hypothesis would reveal that the Mosaic Test functioned no better than chance in identifying individuals on the basis of absence or presence of the behavioral correlates of psychopathology.

Comparisons of both sets of data in Table 4 indicate that the results were relatively unaffected by changes in characteristics of the samples. The slight decreases in values of $\sqrt{\chi^2/N}$ which resulted from lowering cutoff scores may have been due as much to increased sample size as to heightened similarity of samples, and are not the changes which would be anticipated if the Mosaic Test signs were highly sensitive.

When the reaction times of the depressed cluster group were compared with all other cluster groups and the normal and defective diagnostic groups, it was found that the ranges of reaction times among the several groups were such that individuals in every group overlapped with members of the depressed group. Reaction times of several groups were as slow or slower than those of the depressed group and when tests of significance could be conducted, no significant differences were found. The findings remained the same when Wittenborn cutoff scores were lowered to increase the sizes of samples at

Table 4

Occurrence of Hypothesized Mosaic Signs and Values of $\sqrt{\chi^2/N}$ for Upper and Lower Ends of Wittenborn Cluster Groups *

			Lower End		Upper End		
Cluster Group	Sign	References	1. Sign Absent	2. Sign Present	3. Sign Absent	4. Sign Present	$\sqrt{\chi^2/N}$
Acute anxiety	Edge design	12, 14, 16, 17, 19	28 (28)	11 (11)	5 (7)	1 (3)	.75 (.69)
	Corner design	9, 10	28 (38)	1 (1)	6 (10)	0 (0)	.92 .88
	40% or more black in comb. with other Cs. No others preponderant	9, 14	38 (38)	1 (1)	6 (10)	0 (0)	.92 (.88)
Conversion hysteria	Frame design	10, 17, 19	36 (36)	1 (1)	12 (12)	0 (0)	.86 (.86)
	Projections absent or contained	1	32 (32)	5 (5)	11 (11)	1 (1)	.77 (.77)
Manic and depressed states	Emphasis on C without emphasis on F	16, 17	61 (61)	2 (2)	20 (25)	2 (2)	.81 (.78)
Manic state	40% or more red	2, 16, 17	45 (45)	2 (2)	4 (9)	0 (0)	.94 (.9)
	Red projections make up 50% or more of all projections	2, 16, 17	46 (46)	1 (1)	4 (8)	0 (1)	.95 (.88)
Depressed state	Slower reaction time than other groups†	2					
	Emphasis on C without emphasis on F	16, 17	61 (61)	2 (2)	20 (20)	2 (2)	.81 (.81)

* Data in parentheses obtained by lowering cutoffs at upper ends of clusters to obtain larger samples. Where both sets of data are same, it was not necessary (or in some instances, possible) to obtain a larger sample by lowering cutoffs.

† Data summarized in text of article.

Table 4—*Continued**

Cluster Group	Sign	References	Lower End		Upper End		$\sqrt{\chi^2/N}$
			1. Sign Absent	2. Sign Present	3. Sign Absent	4. Sign Present	
	40% or more black plus blue in combination with other Cs. No others preponderant	16, 17	12 (12)	4 (4)	13 (13)	5 (5)	.29 (.29)
	Downward arrow	19	14 (14)	2 (2)	18 (18)	0 (0)	.64 (.64)
Phobic compulsive	Use full 20′ to make coherent unsucc. all-over design	3	50 (50)	2 (2)	7 (14)	0 (0)	.92 (.87)
Schizophrenic excitement	W. tray covered by complex incoherent des. begun within 10″	2	11 (11)	0 (0)	12 (12)	0 (0)	.69 (.69)
Paranoid condition	Small slab	16, 17	38 (38)	0 (0)	4 (7)	0 (0)	.95 (.92)
	Formation of letters of alphabet	—	37 (37)	1 (1)	4 (7)	0 (0)	.94 (.91)
Paranoid schizophrenic	Supersymmetry	2, 16, 17	19 (19)	2 (2)	7 (12)	0 (1)	.79 (.69)
	Repetitive simple des. same shape repeated with variations in color	16	20 (20)	1 (1)	7 (13)	0 (0)	.85 (.77)
Hebephrenic schizophrenic	Simple incoherent design	2, 5, 6, 16, 17	26 (26)	3 (3)	7 (13)	2 (3)	.74 (.63)
	Edging	2	29 (29)	0 (0)	9 (16)	0 (0)	.87 (.8)
	Superimposition of stones	2	28 (28)	1 (1)	7 (13)	2 (3)	.77 (.67)

* Data in parentheses obtained by lowering cutoffs at lower ends of clusters to obtain larger samples. Where both sets of data are came, it was not necessary (or in some instances, possible) to a larger sample by lowering cutoffs.

the most pathological extremes of the cluster groups, and when subjects with reaction times atypical of their groups were not considered.

Lowenfeld's and Wertham's Signs

Examination of Tables 5 and 6 reveal that none of Lowenfeld's or Wertham's signs which had been omitted from the original research plan, or greatly modified for inclusion, functioned as anticipated. Since these signs were not given as hypotheses which had been formulated in advance, cross validation of the findings is indicated.

Discussion

The overwhelmingly negative results justify the conclusion that the Mosaic Test cannot be used in its present form as a tool for differential diagnosis. It may be argued that many of the signs were modified for incorporation into the study and that a molecular rather than a global approach was used. Yet the signs suggested by Lowenfeld and Wertham which were tested separately from the major data also failed to function as anticipated. As regards a global approach to Mosaic analy-

Table 5

Lowenfeld's Signs Not Used as Hypotheses

Sign	References		Exp. Group	Comparison Groups				
			Neurotic	Defect.	Paretic	Schiz.	Maladj.	Normal
Unsuccessful	10	*f*	5	28	26	21	9	14
		z	—	−3.14*	−3.17*	−1.56	−1.53	.63
Slab	10, 17	*f*	0	2	3	3	0	1
		z	—	−1.55	−2.01*	−1.89	0	−.92
			Defect.	Neurotic	Paretic	Schiz.	Maladj.	Normal
No. small, simple designs	9	*f*	8	0	0	4	0	0
		z	—	3.19*	4	1.34	3.19*	4.58*

* Difference significant at 1% level.

Table 6

Wertham's Signs Not Used as Hypotheses

Sign	References		Exp. Group	Comparison Groups				
			Neurotic	Defect.	Paretic	Schiz.	Maladj.	Normal
40% or more black and blue in comb. with other Cs. None other preponderant	16, 17	*f*	3	13	3	15	5	13
		z	—	−1.17	.94	−1.48	−.84	−.28
40% or more red	16	*f*	0	0	0	2	0	6
		z	—	0	0	−1.53	0	−2.3
			Defect.	Neurotic	Paretic	Schiz.	Maladj.	Normal
No. small des., each of one shape	17	*f*	4	0	1	2	0	0
		z	—	.22	1.29	.49	.22	3.17*
No. small, simple, compact, successful	16, 17	*f*	0	0	0	0	0	0
		z	—	0	0	0	0	0
			Schiz.	Neurotic	Defect.	Paretic	Maladj.	Normal
Small slabs	16, 17	*f*	1	0	1	0	0	0
		z	—	1.09	0	1.36	1.09	1.56
Emph. on form without emph. on color	16, 17	*f*	20	5	21	11	5	18
		z	—	1.38	.4	1.54	1.38	2.07
Use of yellow, white, or green alone	16	*f*	1	0	1	0	2	1
		z	—	1.09	0	1.36	−1.38	.3
Black and white used alone	16	*f*	0	0	0	0	0	1
		z	—	0	0	0	0	−1.26

* Difference significant at 1% level.

sis, careful examination of the literature reveals that in spite of statements to the contrary, discrete molecular signs are those most frequently presented as having diagnostic significance.

It seems likely that when the Mosaic Test is used by experts in a clinical setting, many cues which are not directly related to Mosaic performance are used. The source of these cues may be erroneously identified as the Mosaic Test, rather than the interview aspects of the testing situation.

Summary

This paper is a report of an experimental validation study of the Lowenfeld Mosaic Test. The method of group comparisons was used to determine whether there is justification for using the Mosaic Test in its present form as a clinical tool. Mosaic signs were compared to two criteria, psychiatric diagnosis and scores on the Wittenborn Psychiatric Rating Scales.

The results of the study were sweepingly negative and it was concluded that there is no justification for continuing to use the Lowenfeld Mosaic Test in its present form. It was suggested that the discrepancies between clinical impressions and experimental findings may be attributed to failure to recognize that diagnostic cues are obtained from interview aspects of the testing situation, rather than from the Mosaic Test itself.

Received October 6, 1955.

References

1. Colm, Hanna. The value of projective methods in the psychological examination of children: The Mosaic Test in conjunction with the Rorschach and Binet Tests. *J. proj. Tech.*, 1948, **12**, 216–237.
2. Diamond, B. L., & Schmale, H. The Mosaic Test. I. An evaluation of its clinical application. *Amer. J. Orthopsychiat.*, 1944, **14**, 237–250.
3. Dorken, H., Jr. The Mosaic Test: Review. *J. proj. Tech.*, 1952, **16**, 287–296.
4. Eisenhart, C., Hastay, M. W., & Wallis, W. A. *Techniques of statistical analysis*. New York: McGraw-Hill, 1947.
5. Ellenberger, H. The Mosaic Test in use. In Margaret Lowenfeld (Ed.), *On the psychotherapy of children*. London: Heron, 1948. Pp. 62–63.
6. Ellenberger, H. Personal communication. January, 1952.
7. Gurney, Margaret. The use of the Mosaic Test in clinical work with children. In Margaret Lowenfeld (Ed.), *On the psychotherapy of children*. London: Heron, 1948. Pp. 59–61.
8. Levin, M. L. The relationships between construction of form-color-spatial designs and psychopathology. Unpublished doctor's dissertation, Columbia Univer., 1954.
9. Lowenfeld, Margaret. The Mosaic Test. In Margaret Lowenfeld (Ed.), *On the psychotherapy of children*. London: Heron, 1948. Pp. 51–58.
10. Lowenfeld, Margaret. The Mosaic Test. *Amer. J. Orthopsychiat.*, 1949, **19**, 537–550.
11. Lowenfeld, Margaret. Personal communication. November, 1951.
12. Lowenfeld, Margaret. *The Lowenfeld Mosaic Test*. New York: Grune & Stratton, 1952.
13. McCulloch, T. L., & Girdner, J. B. Use of the Lowenfeld Mosaic Test with mental defectives. *Amer. J. ment. Def.*, 1949, **53**, 486–496.
14. Reiman, M. Gertrude. The Mosaic Test: Its applicability and validity. *Amer. J. Orthopsychiat.*, 1950, **20**, 600–615.
15. Shotwell, Anna M., & Lawrence, E. S. Mosaic patterns of institutionalized mental defectives. *Amer. J. ment. Def.*, 1951, **56**, 161–168.
16. Wertham, F. The Mosaic Test: In L. E. Abt & L. Bellak (Eds.), *Projective psychology*. New York: Knopf, 1950. Pp. 230–256.
17. Wertham, F., & Golden, Lili. A differential-diagnostic method of interpreting mosaics and colored block designs. *Amer. J. Psychiat.*, 1941, **98**, 124–131.
18. *Instructions for the Mosaic Test*. New York: Psychological Corp., 1950.
19. *Notes on the nature and use of the Mosaic Test*. London: The Children's Centre, The Institute of Child Psychology, Undated.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

Plausibility and Depth of Interpretation

Seymour Fisher[1]
Walter Reed Army Institute of Research

One of the persistent problems in attempting to study the psychotherapeutic process in a scientifically objective manner lies in the general looseness of psychiatric terminology. Terms such as "emotional insight," "transference," "unconscious hostility"—to mention but a few—are difficult to translate into acceptable operations by which they can be measured. And until satisfactory measurements are available, the *meaning* of a concept (i.e., its empirical interrelations with other variables) cannot be determined to any precise degree.

"Interpretations," both explicit and implicit, are believed to play a major role in many psychotherapies, particularly those which are analytically oriented. Operational definitions of interpretations have been fairly satisfactory (e.g., 5, 11, 12), in the sense that reasonably high interjudge agreement can be obtained in separating "interpretations" from other classes of therapist statements. The question of the "depth" of interpretative statements, however, has received relatively little research attention, although the concept is extremely prevalent in therapeutic literature. Studies by Collier (1, 2) and Harway *et al.* (7) have made preliminary steps toward solving some of the problems in the measurement of depth, but the meaning of the term remains vague and obscure.

The present study had two primary objectives: (*a*) to obtain further information concerning the measurement of depth when the ratings were made under fairly well-controlled conditions; (*b*) to make some introductory exploration into the meaning of the term by comparing "depth" ratings with independent ratings of the same interpretations based upon a concept of "plausibility." It was hypothesized that judgments relating to depth of interpretation are implicitly derived from (and hence should be correlated with) the rater's subjective estimate (i.e., prediction) of how "plausible" the interpretation is to the patient: deep interpretations will be considered more "implausible" than shallow interpretations.

Method

The basic procedure consisted of obtaining ratings of interpretations under a number of different conditions. In order to control certain variables (which will be elaborated at a later point), the present design included the following aspects: (*a*) all interpretations were presented out of context, with the patient's responses omitted; (*b*) the ratings were made from the framework of a single patient whose general personality structure was made known to the raters; and (*c*) the raters were asked to assume that all interpretations were made at a constant time period in therapy.

Statements

Sixty therapist statements were selected from (*a*) recorded protocols, (*b*) published interviews, and (*c*) some were constructed fictitiously. These statements were all considered to be "interpretations" in accord with a definition adapted from Porter (11) and Snyder (12). In choosing the final 60 interpretations, an attempt was made to include samples which might be expected to cover the entire range of "depth."

A fictitious case history and Rorschach summary were composed, and the wording of

---

[1] The writer wishes to acknowledge the technical and statistical assistance of Irvin Rubinstein and Ella J. Wilcox. An abbreviated version of this study was presented at the Eastern Psychological Association, March, 1956.

the interpretations was occasionally modified to "fit" the history. The interpretations were then presented to the judges to be rated within the framework of this hypothetical patient (although the judges were *not* informed that the patient was fictional).

Rating Instructions

Two sets of instructions were drawn up, one referring to "depth," the other to "plausibility." [2]

Depth. The instructions included the following directions to the judge: (*a*) he was requested to rate each of the 60 interpretations on a dimension of depth, "in the usual sense of the term"; no further definition was offered; (*b*) he was informed that all interpretations were made to the same patient, and that an abbreviated case history and Rorschach summary of the patient accompanied the list of interpretations; (*c*) he was asked to assume that all interpretations were given during the first three hours of therapy; (*d*) each interpretation was to be rated on a 7-point scale: a rating of 1 indicating an extremely shallow interpretation, "so shallow that one can hardly consider it to be interpretive in nature"; a rating of 7 indicating the "deepest possible kind of interpretation that could be made to this patient."

Plausibility. These instructions were identical to the "depth" directions, except for the following: (*a*) the term "interpretation" was never used; instead, reference was always made to "therapist statements"; (*b*) the judge was requested to rate each statement for its "plausibility from the patient's point of view: Do these statements seem true or false to this particular patient?"; (*c*) each statement was to be rated on a 7-point scale: a rating of 1 indicating a completely plausible statement; a rating of 7 indicating a completely implausible statement.

Raters

A total of 40 independent judges was used in this study. The judges were assigned to one of four groups depending upon the amount of therapy experience they reported and the set of rating instructions they received. The four groups were composed as follows: (*a*) depth ratings were made by 10 psychiatrists, each of whom had a minimum of 500 hours of therapy experience; (*b*) depth ratings were made by 10 professional psychologists, each having a minimum of 150 hours of therapy experience; (*c*) plausibility ratings were made by a comparable group of 10 professional psychologists; (*d*) plausibility ratings were made by 10 psychology graduate students (first- and second-year students without the M.A. degree), none of whom had any therapy experience.

The experienced judges were deliberately selected from different geographical regions to minimize personal friendship and similar training among raters.

It will be noted that no psychiatrists were asked to rate for "plausibility," since it was considered highly probable that despite rating instructions they would most likely still be rating for depth. Similarly, no inexperienced graduate students were asked to rate for "depth," since one would not expect them to have any notion as to the meaning of the term.

In Table 1 the salient characteristics of the four groups of raters are summarized. It is clear that the psychiatrists, psychologists, and graduate students differ primarily in respect to therapeutic experience, although in addition the graduate students tend to be somewhat younger than the professionals. Thus, it should be kept in mind that although the groups will hereafter be referred to as "psychiatrists," "psychologists," and "graduate students," differences among these groups should not be attributed too literally to the different roles; it is believed that, essentially,

Table 1
Median Age, Hours of Therapy Experience, and Hours of Personal Therapy of the Four Groups ($N = 10$) of Raters

Raters	Instructions	Age	Hours of therapy experience	Hours of personal therapy
Psychiatrists	depth	32.0	2,275.0	239.0
Psychologists	depth	32.5	490.0	25.0
Psychologists	plausibility	33.0	430.0	0.0
Grad. students	plausibility	25.0	0.0	0.0

---

[2] Copies of the rating instructions, items, and case history along with complete tables of intercorrelation matrices for the four groups of raters and scatter diagrams of the correlations in Table 2 have been deposited with the American Documentation Institute. Order Document No. 4813 from ADI Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.75 for microfilm or $2.50 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

these groups represent points along a dimension of therapy-experience.

Results

Depth Measurements

If there is any common meaning attached to the term "depth" one would expect to find significant intercorrelations among the judges who rated the 60 interpretations. The median r ($N = 45$) for the 10 psychiatrists[3] was .69 ($Q_1 = .61$, $Q_3 = .74$), and for the 10 psychologists .78 ($Q_1 = .72$, $Q_3 = .83$). Obviously, then, the raters were employing common criteria in judging the "depth" of the items. Since the distribution of intercorrelations within the psychiatrists and psychologists was relatively homogeneous and symmetrical, it was possible to compare the ratings of the two groups: for each of the 60 items, the mean ratings of the 10 psychologists were correlated with the mean ratings for the 10 psychiatrists. The relationship was a marked linear one, $r = .96$. The scatter diagram also suggested a difference in the means of the two distributions. The 60 observations, however, were not independent (the same judges were involved), precluding a direct statistical test of significance. Therefore, the mean of each rater's 60 ratings was used as a total score, and a t test was computed between the two groups of raters. The mean rating for the psychiatrists was 3.67 ($SD = .16$); for the psychologists 4.15 ($SD = .26$). The t of 4.67 is significant at the .001 level for 18 df, indicating that psychologists tend to attribute greater depth to interpretations than do psychiatrists.

Plausibility Measurements

Similar analyses were performed on the plausibility ratings done by 10 psychologists and 10 early graduate students. The median r for the psychologists was .63 ($Q_1 = .56$, $Q_3 = .68$), and for the naive raters .68 ($Q_1 = .64$, $Q_3 = .74$), signifying that these plausibility ratings can similarly be done with a considerable amount of interrater agreement. The median r's for plausibility tend to be somewhat lower than the depth ratings, although the interdependence among the coefficients does not permit a statistical test of the significance of the differences. It may well be that depth ratings can be made more "reliably" than plausibility ratings, although it is probably more parsimonious to assume (until replication or additional evidence to the contrary) that these are chance differences.

When the 60 mean ratings by the psychologists were plotted against the mean ratings of the graduate students, the Pearson r was .96. Here, too, there was the suggestion that the less experienced graduate students were rating items as more implausible than the psychologists' ratings: the mean rating for the psychologists was 3.88 ($SD = .40$), for the graduate students 4.21 ($SD = .33$). The t of 1.9 does not quite reach the .05 level of significance (two-tailed test) for 18 df.

Depth and Plausibility

If the plausibility ratings are compared to the depth ratings, the relationship should indicate to what extent these two concepts are measuring the same or different things. Hence, the mean plausibility ratings of the psychologists were plotted against the mean depth ratings by psychologists. An r of .87 was obtained, suggesting that ratings based upon plausibility estimates are essentially the same as depth ratings. Between these two groups of raters, where experience was equated, the mean ratings on the scales (3.88 for plausibility, 4.15 for depth) did not differ significantly ($t = 1.7$).

When the plausibility ratings by early graduate students were plotted against the depth ratings by psychiatrists, a similar relationship appeared: $r = .86$.

Table 2 summarizes the intercorrelations among the four *groups* of raters mentioned above, including also the correlation between graduate students' plausibility ratings and psychologists' depth ratings ($r = .88$), and between psychologists' plausibility ratings and psychiatrists' depth ratings ($r = .84$).

---

[3] One psychiatrist, who met the criteria for inclusion in the sample, but whose ratings were conspicuously deviant from his colleagues', was (after much consideration and debate) dropped from the analysis. Although he claimed that he had understood the instructions, it appeared that his judgments were decidedly atypical. His ratings correlated with his ten colleagues respectively as follows: − .45, − .25, − .32, − .36, − .37, − .43, − .37, − .26, − .40, − .42.

Table 2

Intercorrelation Matrix of Mean Ratings ($N = 60$) Among Four Groups of Raters

	Psychiatrists depth	Psychologists depth	Psychologists plausibility	Grad. students plausibility
Psychiatrists depth	—	.96	.84	.86
Psychologists depth		—	.87	.88
Psychologists plausibility			—	.96

Discussion

Measurement of Depth

In devising this study it seemed desirable to consider and attempt to control a number of factors which might be expected to affect the ratings. Three such factors appeared particularly important.

1. A rather common way of estimating depth clinically seems to be based upon the patient's subsequent reaction to an interpretation. If the patient strenuously rejects an interpretation, it must perforce be deep; an avidly accepted interpretation suggests shallowness. This view is implied, for example, by Fenichel (3) and is stated explicitly by Hilgard (8, p. 26). Nevertheless, this approach has certain limitations. Foremost is the consideration that, for practical purposes, a therapist would be unable to decide in advance whether he should withhold a particular interpretation because it might be too "deep" at that time. Furthermore, since it appears that some therapists are highly astute in selecting propitious moments to offer interpretations, it would seem that depth can be estimated prior to, and independent of, knowledge of the patient's postinterpretive responses.

2. An adequate scale of depth cannot, however, be independent of the patient to whom the interpretations are offered. The very same interpretation may be "deep" for one patient while quite "shallow" for another. Certainly, a sexual interpretation to a psychiatric resident analysand may have different depth than the identical interpretation given to an illiterate farmhand. This point is frequently neglected in loose discussions on the meaning and measurement of depth. For example, Collier's scale (2) is constructed purely on the basis of the type of therapist behavior, quite independent of any specific patient. Yet his scale is derived from a concept of "degree of uncovering" (i.e., the more the interpretation "uncovers," the greater is its depth) which clearly implies the dependence of depth upon a specific personality structure: how much depth does an "uncovering" interpretation have when a patient, prior to the interpretation, is already aware of the material which the therapist is "uncovering"?

3. Since depth of interpretation is often conceived in relation to the degree of "awareness" on the patient's part, one might expect that depth ratings will vary as a function of the time in therapy when an interpretation occurred. As therapy progresses, it is generally claimed that patients tend to acquire increased awareness. Hence, an interpretation given late in therapy might not be as deep as if it were given very early. It is also possible, however, that a late interpretation be no deeper than the same interpretation given earlier (i.e., if—relative to the content of the interpretation—the patient's degree of "awareness" remains unchanged). The mere knowledge that an interpretation occurred in the two-hundred-fiftieth hour rather than the first hour is conceptually irrelevant to how deep it is. There is nothing inherent in the variable of time per se as a determinant of depth. The substantive point concerns knowledge of the patient's personuality at the two-hundred-fiftieth and first hours.

Under these three conditions, the results of the present study show that interpretive levels can be estimated with a high degree of reliability by using either "depth" or "plausibility" ratings. While mean plausibility ratings may perhaps be slightly more variable (*SD* of .40 and .33 compared with .16 and .26), they have the distinct advantage of being available for use with comparatively inexperienced judges. For a series of interpretations, averaging the ratings of a small number of therapeutically naive graduate students will give a fairly accurate estimate of the mean depth rating obtained from experienced psychologists ($r = .88$) or even more experienced psychiatrists ($r = .86$). Thus, in the tedious and time-consuming analysis of therapeutic protocols, the need for professional raters might be eliminated, substituting instead reasonably bright "technicians." It should be noted, however, that less experienced judges tend to shift their ratings toward the deeper (more implausible) end of the scale, a finding originally observed by Harway *et al.* (7). In certain circumstances this additional variable would have to be taken into account.

Of course, in one sense, the correlations obtained in this study are misleadingly high, inasmuch as items sampling a broad range of depth were used. In actual practice, raters would more often be confronted with a truncated sample in any one case (limited by the theoretical scope and technique of the therapist), and consequently lower correlations would be expected. On the other hand, the relatively narrow distributions of interjudge correlations are indicative of the satisfactory reliabilities to be obtained when specific controls are placed on the measurements.

Meaning of Depth

In the present study, none of the judges was given any definition of the term "depth" when making his ratings. Nevertheless, the mean rating of the psychologists correlates .96 with the mean rating of the psychiatrists. There seems little doubt, then, that experienced therapists of various orientations are employing some common cues in determining depth. This does not mean, however, that these therapists share a common, verbalized *meaning* for the concept. Many of the judges spontaneously offered comments on their ratings. The most significant aspect of these comments concerned the variability of the criteria which *the raters believed they were utilizing*. Among the 20 judges who made depth ratings, a variety of definitions was offered:

> I tend to equate depth with the impact the statement might be assumed to have on the patient, and with the lack of possibility of verification of the statement by observation of the patient's behavior.
>
> I used such criteria as the directness of the vocabulary in the interpretation as related to deep unconscious drives; those that were related to instinctual urges were deeper, those related to ego defenses more shallow.
>
> . . . ratings certainly involved at least a couple of dimensions besides depth—psychoanalyticness, range of generalization, centralization, centrality of problem involved, etc.
>
> . . . depth is in itself an ambiguous concept, since it can be defined from several points of view, e.g., developmental-genetic, resistance of subject to accepting it, etc.
>
> I found myself giving ratings toward the "deep" end when: (1) interpretation was of material defended against, not of defense itself; (2) when interpretation was historical; (3) when interpretation tied past and present; (4) when interpretation was of symbolic material rather than of overt behavior (e.g., dreams, phantasies).
>
> I sense that I tend to regard the more symbolic comments as "deeper." What deeper means is a little obscure. "Nearer manifest content" and "assumed latent content" would seem to be approximates of "shallow" and "deep."

From the preceding list, it would appear that—on a verbal level at least—different therapists would have difficulty in agreeing among themselves as to the meaning of the term "depth." Some therapists, in fact, even felt that they lacked consistency in their ratings (e.g., "I'm certain if I were to do it a second time my reliability would be very low"). Yet the generally high intercorrelations among the depth raters implies that this is not really the case, and that there is a considerable amount of reliability in the cues used for rating the present sample. (One psychiatrist was kind enough to repeat his ratings after a two-day interval: test-retest r was .94.)

The precise nature of these cues, of course, remains unknown. The data do, however, suggest that inferences in regard to depth are possibly based upon cues similar to the ones used in drawing plausibility inferences. The cues used by an inexperienced rater in inferring the plausibility of a statement may or may not be the very same cues an experienced therapist uses in estimating depth. In either case, however, there is the interesting implication that at least some of the cues believed to be important by professional therapists may be irrelevant: specifically, that class of cues associated with a therapist's theoretical and professional training. Within the limits of the present results, it seems that advanced training in therapy and psychodynamics is not prerequisite for a knowledge of what is "deep" or "shallow"; individuals without such training are capable of arriving at essentially similar conclusions.[4]

Depth and Plausibility

In view of the strong relationship between depth and plausibility ratings, it is tempting to intimate that when a therapist refers to the depth of an interpretation, he is doing nothing more than making an estimate of how implausible his statement would appear to this particular patient at this particular time. Clearly, the data from this limited study do not provide sufficient support for such a contention. But even in the absence of adequate empirical verification, this oversimplified notion has a number of features which might be elaborated.

1. If we equate depth with perceived plausibility, it becomes possible to understand why many therapists verbalize their definition of depth in terms of the patient's response to an interpretation. It is much easier to infer how plausible a statement *was* (using the patient's subsequent behavior as additional information) than to estimate how plausible the statement *will be*. Without such information, however, it is still possible to make predictions of the patient's subsequent reactions based upon knowledge of his previous behavior. Of incidental interest is the fact that other investigators (5, 10) have reported that therapists are no better than laymen in making these predictions.

2. While the concept of depth assumes truth-content of an interpretation, this limitation disappears when a dimension of "perceived truth" or plausibility is used. As Festinger (4) has observed, many opinions have no inherent truth value in respect to the physical world. This seems particularly applicable to the content of therapeutic interpretations: whether an interpretation is deemed "true" or not generally depends upon the training of the therapist, with an Adlerian disputing the "truth" of a Freudian interpretation, etc. (for some interpretations referring to the effects of repressed hostility and sexual impulses, one suspects that the *Zeitgeist* also plays a role in determining their alleged truth value). Suppose a therapist says, "Your anxiety stems from the devil, an evil spirit who lives in your stomach and makes you quiver inside and outside every time you become angry." Is this interpretation deep? Of a number of therapists who were asked this informally, many replied that the question had no meaning, that the "interpretation" could be considered neither deep nor shallow because it is patently absurd (i.e., obviously "untrue").[5] Nevertheless, this may not be absurd to a particular kind of patient. For one who believes in the reality of devils residing in the viscera, the therapist's statement may be perceived as being true (i.e., plausible), and conceivably could have a significant effect upon the therapeutic process. The concept of plausibility permits evaluation of *any* therapist interpretation, irrespective of the truth or falsity of its content.

3. From the working hypothesis that depth and plausibility refer to the same underlying construct, it is possible to attempt to integrate recent experimental studies in social psychology with the study of the psychotherapeutic process. Depth can be considered

---

[4] This is not meant to deny the possibility that the professional therapist is more sensitive to relevant cues as a result of his experience. In the present study all possible cues were neatly presented to the judges in the case history and Rorschach summary, thus minimizing the importance of the rater's perceptive abilities. In a situation where the number of cues is virtually infinite (e.g., a live doctor-patient interaction or even the evaluation of a complete series of protocols), the inexperienced individual might be seriously handicapped in deciding how plausible a statement is to a particular patient.

[5] But if a therapist in 1956 were to tell a patient, in more simple language perhaps, that his anxiety was due to the practice of coitus interruptus (4, p. 187), would this be considered as "true"? It is interesting to note that virtually all common definitions of an "interpretation" tacitly assume the empirical validity of the causal relationship expressed by the therapist. Whether one talks about "uncovering," "unconscious" (i.e., patient being "unaware"), or the like, the frame of reference always seems to be one in which it is hoped that the patient will ultimately come to recognize the "truth" of what the therapist is implicitly or explicitly communicating. Perhaps it is sufficient for the patient to *accept* (in the sense of internalizing a new basic belief) what the therapist is suggesting.

as a form of "distance," a concept that is receiving some current attention and one which can be controlled and manipulated systematically in a laboratory setting. By "distance" is meant the discrepancy between two persons' opinions. That is, if *P* (a prestige source) expresses an opinion quite similar to that held by *S* (subject), the distance is small; if *P* expresses a judgment markedly different from *S*'s belief, the distance is large. It is self-evident that, other things being equal, plausibility is a necessary monotonic correlative of distance, larger distances being more implausible than smaller distances. Now an interpretation is a particular type of judgment by a therapist, one which can be conceptualized as lying along a dimension of distance. A "deep" (implausible) interpretation is one with a large distance; a "shallow" (plausible) interpretation has little distance. A therapist statement which has zero distance is *not* an interpretation (e.g., patient says "I hate my father"; therapist then says "You hate your father"). Thus a new and unified working definition of interpretation is suggested, one which *includes* in its definition the dimension of depth and at the same time eliminates the need for any assumptions regarding truth value:

An interpretation is an implicit or explicit judgment about the patient's motivational and emotional behavior where the distance is greater than zero; the magnitude of the distance is a measure of the interpretation's depth.

Implications

If one is willing to entertain the possibility that psychotherapy is an instance of a two-person interaction series where one individual (therapist) attempts to modify the belief-systems of another individual (patient) by "social influence," then some recent laboratory findings might have bearing upon the study of therapy. For example, in a two-person group it has been shown that distance (and consequently plausibility) is a significant variable in determining the degree to which *S*'s verbal judgment will conform to *P*'s: in general, the larger the distance the less *S* will conform, the function being negatively accelerated (9). From this generalization, it should follow that, other things being equal, deep interpretations should be rejected by the patient more frequently than shallow interpretations. It has also been found that when continued pressure is placed upon *S* over a series of interactions with a prestige figure, under certain conditions *S* will subsequently show more conformity to judgments by *P* than ever would have been expected had *P* made these judgments early in the interaction series (6). In other words, judgments which would have been extremely implausible if given at the beginning of interaction between *P* and *S* tend to become considerably more plausible after certain kinds of interaction have occurred over a period of time. These results may possibly offer a clue as to how some patients, following long-term psychotherapy, have acquired beliefs from their therapist (e.g., "castration fear is the basis of my difficulties") which in all probability would have been rejected if the therapist had introduced these interpretations in the initial phases of therapy.

As more research findings emerge concerning effects of distance in continuous interactions, it would not be too surprising to find that the role played by interpretations in the nebulous process of psychotherapy is considerably illuminated.

Summary and Conclusions

Four groups of judges, three of which differed in therapeutic experience, were asked to rate a series of 60 psychotherapeutic interpretations. Two groups rated for "depth"; two groups rated for "plausibility." High intragroup agreement was obtained on both depth and plausibility scales. When mean plausibility ratings were correlated with mean depth ratings, a sufficiently high relationship was found to suggest that the two scales may be measuring the same underlying construct. Plausibility ratings of therapeutically naive graduate students correlated .88 with depth ratings from experienced psychologists and .86 with psychiatrists' depth ratings, implying that advanced training in therapy and psychodynamics may not be prerequisite for a knowledge of what is "deep" or "shallow."

It was also noted that less experienced raters tend to judge interpretations as more deep (and implausible) than experienced therapists. The possible advantages and implications of conceptualizing depth in terms of plausibility are discussed.

Received September 13, 1955.

References

1. Collier, R. M. A basis for integration rather than fragmentation in psychotherapy. *J. consult. Psychol.*, 1950, **14**, 199–205.
2. Collier, R. M. A scale for rating the responses of the psychotherapist. *J. consult. Psychol.*, 1953, **17**, 321–326.
3. Fenichel, O. *The psychoanalytic theory of neurosis.* New York: Norton, 1945.
4. Festinger, L. A theory of social comparison processes. *Hum. Relat*, 1954, **7**, 117–140.
5. Fiedler, F. E. Quantitative studies on the role of therapists' feelings toward their patients. In O. H. Mowrer (Ed.), *Psychotherapy: theory and research.* New York: Ronald, 1953. Pp. 296–315.
6. Fisher, S., Rubinstein, I., & Freeman, R. W. Inter-trial effects of immediate self-committal in a continuous social influence situation. *J. abnorm. soc. Psychol.*, 1956, **52**, 200–207.
7. Harway, N. I., Dittman, A. T., Raush, H. L., Bordin, E. S., & Rigler, D. The measurement of depth of interpretation. *J. consult. Psychol.*, 1955, **19**, 247–253.
8. Hilgard, E. Experimental approaches to psychoanalysis. In E. Pumpian-Mindlin (Ed.), *Psychoanalysis as science.* Stanford, Calif.: Stanford Univer. Press, 1952. Pp. 3–45.
9. Lubin, A., & Fisher, S. Distance as a determinant of influence in a two-person continuous interaction situation. Paper read at East. Psychol. Ass., Atlantic City, March, 1956.
10. Luft, J. Implicit hypotheses and clinical predictions. *J. abnorm. soc. Psychol.*, 1950, **45**, 756–759.
11. Porter, E. H., Jr. The development and evaluation of a measure of counseling interview procedures. *Educ. psychol. Measmt*, 1943, **3**, 105–126; 215–238.
12. Snyder, W. U. (Ed.). *Group report of a program of research in psychotherapy.* State College: Pennsylvania State College, 1953.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

Personality Patterns of Neurotic Adults in Psychotherapy

Maurice Lorr and Eli A. Rubinstein[1]
The Veterans Administration and Catholic University of America

There is wide agreement that present-day methods for evaluating outcomes of psychotherapy are unsatisfactory. There are no generally accepted measures of morbidity or of change and few that have gained even a partial acceptance from research workers. What is needed is a more objective system of evaluation than is currently available and a set of measures descriptive of the principal dimensions of personality and mental illness. The study reported here is one part of a program designed to develop criteria of mental illness and to isolate some of its more important dimensions.

In a recent analysis by the authors (3) the correlations between 58 ratings of a group of 184 veterans in psychotherapy were found to be well accounted for by ten personality and symptom patterns. Seven of these patterns were found to bear varying degrees of resemblance to factors identified in earlier studies. However, it seemed highly desirable to establish more firmly the existence of these ten parameters, of which several had not been identified previously. Clearly if these factors could be isolated repeatedly in patient samples differing widely in composition and with varying combinations of reference variables, they would be acceptable at a much higher level of confidence.

With this goal in mind the present study was designed to test whether ten personality and symptom patterns recently identified could be confirmed and clarified in a second group of patients. The initial study was based on neurotic and psychotic patients who had been in treatment on the average for three months when rated. In this analysis the patients included were restricted to neurotics who had been in treatment for three or four sessions when rated. The rating schedule used in the original study was also modified. The descriptive scales used for rating patients in the initial analysis were converted into graphic form. Scale titles were removed and scales were presented in a random order to the raters. The question was whether the same ten parameters would be isolated again despite these changes in composition of the sample and alteration in the format of the rating schedule.

The study was structured around ten hypothesized constructs. The underlying patterns or syndromes included for experimental test were the following: (*a*) lack of emotional restraint vs. inhibition of emotionality, (*b*) degree of distortion of reality in thinking and perception, (*c*) anxious tension vs. relaxed comfort, (*d*) sense of personal adequacy vs. its lack, (*e*) dependent immaturity vs. independent immaturity, (*f*) conscientiousness vs. its lack, (*g*) obsessive-phobic reaction, (*h*) gastrointestinal reaction, (*i*) cardio-respiratory reaction, (*j*) conflict between sex impulses and moral standards.

Included also were five new scales descriptive of alcoholic addiction, defensiveness, impulsiveness vs. deliberateness, avoidance of attention vs. self dramatization, and acceptance vs. resistance to regulation by others. These were constructed to further clarify factors already identified.

Procedure

The sample consisted of 215 nonpsychotic male World War II and Korean veterans with service-incurred psychiatric disabilities. Eleven Veterans Administration mental hygiene clinics ranging, geographically, from

---

[1] From the Veterans Benefits Office, Washington, D. C.

Los Angeles to Newark collaborated in the study.[2] All patients were newly accepted for treatment and 90 per cent had never before received any formal psychotherapy. The median age was 32 and the median grade completed was 11.

Each patient was rated once by his own therapist on each of the 61 scales of the MSRPP (4) after the third or fourth interview. The raters were predominantly staff therapists (psychiatrists, clinical psychologists, and social workers), but a few were advanced psychology trainees and psychiatric residents. Raters in each clinic participated in several training sessions and were provided ample opportunity to discuss and compare their practice ratings of patients and colleagues.

The rating schedule (MSRPP) consisted of 61 unlabeled, randomly ordered, four- and six-point graphic rating scales. A detailed rating guide was attached as a preface to each rating form. Independent studies of interrater agreement indicated that the average product-moment correlations between raters following two or three training sessions was about .75.

Fifty of the 61 scales were selected for the factor analysis. Of the 11 scales omitted most were dropped because of lack of relevance to the study, a few because they were too often left unrated, and the remainder were eliminated if the symptom (e.g., skin disturbances) was rarely noted. Product-moment correlations between the 50 scales were obtained by IBM procedures. The variable correlations were then arranged into ten clusters on the basis of the factors hypothesized. The ten clusters and their defining variables were as follows: Cluster *A*, variables 1 to 6; Cluster *B*, variables 7 to 13; Cluster *C*, variables 19 to 24; Cluster *D*, variables 25 to 29; Cluster *E*, variables 30 to 34; Cluster *F*, variables 35 to 39; Cluster *G*, variables 14 to 18; Cluster *H*, variables 40 to 44; Cluster *I*, variables 44 to 46; Cluster *J*, variables 47 to 50. See Table 1.

From the 50 × 50 clustered correlation matrix, 10 factors were extracted simultaneously by the multiple group method with centroid estimates of the communalities in the diagonal cells. No evidence of additional factors was observable in the residuals. The obtained oblique matrix was orthogonalized and then rotated blind by the single-plane method to oblique simple structure. The structure was sufficiently well defined by the initial cluster arrangement that on the average only two rotations were required to locate a plane. The oblique rotated factor matrix V is given in Table 2.[3]

The First-Order Factors

Only those variables whose correlations with a factor are .30 or higher are used in the interpretations of the first-order factors.

Factor *A* defines a clear-cut bipolar parameter of emotional responsiveness identical with the *A* previously reported. A patient scoring high on *A* is relatively unrestrained in his behavior, moody, overreactive emotionally, inclined to dramatize himself, impulsive and overt in his expressions of hostility. In addition he is likely to be cheerful in mood and enthusiastic in his interests. A low-scoring patient is relatively inhibited and restrained in his emotional behavior, unable to express anger, hate, or other feelings openly, and inclined to avoid drawing attention to himself. Clinicians are likely to recognize the extremes of this pattern. It is common practice to categorize some patients as overcontrolled and unresponsive emotionally and others as lacking in restraint and control.

Factor *B* appears to be measuring a parameter descriptive of a hostile, suspicious rebelliousness against adults in authority. At the maladjusted end the patient is hostile and suspicious of others, resentful of authority, defensive in attitude, inclined to blame others for his difficulties, unlikely to give way in conflicts, and is resistive to regulation by others. In addition he is markedly lacking in concern for the welfare of others. Factor *B* resembles the *B* of the earlier study with respect to maladjustment in interpersonal relations but it lacks the gross distortion of thinking and perception of the latter. In the initial study, reality distortion, No. 14, and ideas of reference, No. 18 correlated with *B*. In the

---

[2] The continued cooperation of the respective clinic chiefs, chief clinical psychologists, and station investigators from the Veterans Administration mental hygiene clinics in Baltimore, Brooklyn, Chicago, Denver, Detroit, Los Angeles, Newark, Pittsburgh, St. Paul, San Francisco, and Washington, D. C., is most gratefully acknowledged. The authors also acknowledge the invaluable assistance of Elizabeth Turk with the statistical computations.

[3] The following additional tables have been deposited with the American Documentation Institute: Table A, Orthogonal Factor Matrix F of Correlations of 50 Scales with Each Factor; Table B, Transformation Matrix; Table C, Correlations Between Primary Factors; Table D, The Second Order Factor Matrix F. Order Document No. 4850 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting, in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Table 1

Brief Descriptions of Both Ends of the 50 Scale Variables Factored

1	unrestrained feeling	restrained feeling
2	expresses hostility overtly	conceals hostility
3	emotional over-response	emotional under-response
4	frequent mood changes	no mood changes
5	highly impulsive	highly deliberate
6	dramatizes self	shrinks from attention
7	blames self for difficulties	blames others for difficulties
8	bears little hostility	bears much hostility
9	perceives world as friendly	perceives world as hostile
10	acceptant of authority	resentful of authority
11	rarely defensive	usually defensive
12	open and trusting	usually suspicious
13	accepting of regulation	resistive to regulation
14	no distortions of reality	marked distortions of reality
15	no insistent thoughts	frequent recurring thoughts
16	attentive to outside matters	self preoccupied
17	no morbid fears	disrupted by morbid fears
18	no unjustified suspicion	has ideas of reference
19	relaxed	tense
20	rare sleep difficulty	frequent sleep difficulty
21	rarely irritable	frequently irritable
22	elated	depressed
23	untroubled	disrupted by anxiety
24	alcohol not a problem	alcohol disrupting behavior
25	rarely gives way to others	usually gives way to others
26	unaffected by approval need	much affected by approval need
27	seldom feels inadequate	usually feels inadequate
28	strong belief in self	weak belief in self
29	untroubled by guilt	much troubled by guilt
30	neglects responsibility	over-accepting of responsibility
31	dependent on others	wholly self reliant
32	weak interests	strong, active interests
33	low achievement motivation	high achievement motivation
34	low standards	high standards
35	weak impulse control	strong impulse control
36	careless with tasks	painstaking with tasks
37	little concern for orderliness	strong concern for orderliness
38	no concern for anyone	strong concern for others
39	never considers tomorrow	much consideration of the future
40	no headache complaints	disturbing headaches
41	energetic	tired and worn out
42	no gastric symptoms	severe gastric symptoms
43	no intestinal or bowel symptoms	severe intestinal or bowel symptoms
44	no body concern	extreme body concern
45	no respiratory symptoms	severe respiratory symptoms
46	no cardiovascular symptoms	severe cardiovascular symptoms
47	no sex conflict	disturbed by sex conflict
48	no compulsive acts	disturbed by compulsive acts
49	no concern over homosexual tendencies	concern over homosexual tendencies
50	no concern over masturbation	guilt over masturbation

present study these two variables have their principal projections on *G*. These differences are possibly due to the absence of patients with obvious psychotic symptomatology in the present sample.

The third Factor *C* appears to be identical with the *C* of the earlier study except for the presence of a depressive element. A manifest tension, depressive in character, appears to underlie the factor. The maladjusted end of *C* is defined by signs of manifest tension, an apprehensiveness not called for by circum-

Bureau Ednl. Psy. Research
DAVID HA... TRAINING COLLEGE
Dated ....
Acc. No ....

Table 2

Rotated Oblique Factor Matrix V
(Decimal points are omitted.)

Scale No.	Factors									
	A	B	C	D	E	F	G	H	I	J
1	69	06	04	07	06	02	02	13	07	−05
2	45	−33	−08	34	19	−25	00	06	11	−08
3	62	−04	02	−04	−11	−07	−03	−01	−12	06
4	39	−08	−02	08	−05	−05	−03	−05	−06	−13
5	26	−23	−20	16	−04	45	−06	03	03	−01
6	48	−09	14	−13	−18	25	10	−01	−07	05
7	−17	47	−21	−12	−17	25	−03	02	00	−22
8	−04	70	07	−03	−05	19	03	−05	−09	01
9	05	51	15	33	02	00	01	13	16	−06
10	−01	68	−02	08	16	−04	00	08	11	04
11	−02	53	−01	13	09	16	−18	06	04	−04
12	22	57	20	01	06	−01	18	16	−11	−12
13	−18	55	04	−08	13	−17	−02	03	−01	02
14	−07	18	−01	03	−08	01	50	04	00	−02
15	02	03	−02	00	08	−03	56	−03	−02	01
16	10	19	14	00	−15	−06	34	01	07	−06
17	−03	00	−11	−03	03	15	40	03	15	19
18	07	22	05	−05	−09	07	41	04	−06	−08
19	−12	04	43	17	−02	−06	00	02	−02	06
20	01	−09	31	−04	−11	05	06	31	−02	15
21	−25	33	35	−14	−10	06	−03	04	−01	16
22	30	18	33	−07	−23	27	19	02	−19	−13
23	01	02	42	18	05	07	12	−02	25	−02
24	−08	06	24	−08	−02	−25	09	−04	09	15
25	17	−37	−06	42	−25	06	−12	−05	09	09
26	−25	−19	−15	35	16	22	−11	04	23	06
27	03	−01	−01	65	00	−02	01	01	00	00
28	01	−07	03	51	−39	05	04	−21	10	−11
29	12	−14	05	21	03	16	−07	−02	02	29
30	−02	−17	−08	−12	33	40	−05	05	01	−09
31	05	06	11	−28	61	−13	−05	14	−23	01
32	−37	−05	−18	−02	63	−04	−08	12	−19	00
33	−05	−02	−05	−02	72	06	01	05	00	02
34	02	01	03	21	70	−03	01	−05	−03	−03
35	26	−29	−13	13	25	16	−06	03	06	−08
36	08	−03	02	−20	26	48	02	−03	−06	−02
37	−01	01	01	01	−02	64	−02	05	01	−01
38	−05	−48	−03	02	27	13	−05	00	06	02
39	19	−09	23	−03	53	11	−08	−05	02	10
40	13	−06	34	−03	−09	−08	−17	17	24	04
41	19	00	01	19	−39	11	15	29	15	−13
42	−03	−13	−04	03	−03	05	−01	67	−05	−05
43	09	01	06	−02	15	−08	06	64	05	06
44	04	−01	−03	03	−19	29	00	55	18	06
45	−10	−07	−15	10	06	−03	02	−09	56	04
46	03	02	00	−05	−04	04	−02	11	65	−06
47	10	12	00	−08	−09	02	20	−03	−03	48
48	−02	05	−03	−12	12	−04	23	03	−02	34
49	06	17	−04	−04	02	08	08	−03	15	38
50	−01	−07	00	03	07	−03	−02	04	06	57

stances, irritability, indications of a depressive mood, and difficulties with sleep. In clinical practice Factor *C* would probably be regarded as an anxiety reaction.

The fourth Factor *D* is descriptive of the sense of personal adequacy in the patient. At the maladjusted end of the scale the patient has a marked tendency to feel inadequate, is lacking in a belief in himself and his powers, is inclined to give way and defer to others in conflicts, and is much affected in his behavior by a need to secure and maintain their approval. Except for the absence of significant projection from guilt feelings, No. 36, *D* is the same as the Factor *D* isolated in the earlier study.

E is a bipolar factor descriptive of drive toward long-term goals. At one end of the continuum is the patient whose motivation for long-term goals such as achievement, acceptance, recognition, or power is low, and whose standards of adequacy of performance, in areas he considers important, are also low. His interests are weak, and his energy is low and he is lacking in belief in himself. He seems never to consider tomorrow, and is inclined to depend on others for the initiation and direction of his activities. The opposite end of the continuum describes the individual whose activities are much determined by long-term goals. His interests are strong, his energy is high, and he is inclined to be self-confident and self-reliant. In the initial study a corresponding Factor *E* of independent maturity was identified. This factor contained elements of what might be described as responsible maturity and goal-directed control or persistence. In the present study the goal-direction and independence are more prominent than tendencies toward acceptance of adult responsibility.

The bipolar Factor *F* is defined at one end by a strong concern for system, orderliness, or routine in activities and surroundings, and by the careful and painstaking performance of tasks undertaken. An inclination to be deliberate and reflective prior to action and a tendency to accept more than one's share of responsibility and obligation is also characteristic of *F*. The opposite end of *F* is characterized by a careless indifference to the performance of tasks, a lack of concern about adherence to accustomed patterns, impulsiveness and a tendency to avoid or neglect responsibilities. Patients scoring high on the conscientious pole of *F* correspond closely to the compulsive personality delineated in the American Psychiatric Association's manual of mental disorders (1). Such patients are said to be characterized by chronic, excessive, or obsessive concern with adherence to standards of conscience or conformity. A character trait of conscientiousness vs. careless indifference appears to be represented in *F*. Essentially the same factor was isolated in the initial study.

In the initial study it was stated that Factor *G* appears to be defining an obsessive-compulsive-phobic reaction. However, it was felt that this inference required cross checking. In the present study the patient scoring high on *G* tends to be preoccupied with insistent, recurring, useless thoughts, tends to distort reality in his perception and thinking, is disturbed by morbid fears, and is inclined to believe on the basis of slight evidence that people talk about or refer to him. As might be expected he is also described as much preoccupied with himself. The hypothesis of an obsessive-compulsive reaction is plausible but the ideas of reference seem out of place. An alternative hypothesis is that *G* represents one aspect of an incipient or latent schizophrenic tendency. Obvious psychotics exhibiting delusions, hallucinations, and grosser forms of reality distortion and falsification were not included in the present study. However, no rigorous effort was made to exclude patients who might show indications of schizophrenic tendencies, since the data were collected for another purpose. The prominence of distortions of thinking and perception and the extent of disruption of normal activities suggest that this parameter represents a more profound disturbance than an obsessive-phobic reaction.

The scales defining Factor *H* describe a gastrointestinal reaction. The high-scoring patient is likely to present symptoms of stomach, intestinal, and bowel disorders, and accordingly is likely to be much preoccupied with his health and the functioning of his bodily organs. Except for the absence of headaches, Factor *H* is identical with the corresponding factor previously identified.

Factor *I* is defined only by the presence of symptoms of respiratory disturbance (asthma, hay fever) and of cardiovascular symptoms (tachycardia, hypertension). As in the earlier study *I* represents a cardiorespiratory reaction. It is possible that if the specific syndromes included under these two scales (such as asthma and hypertension) were individually scaled, Factor *I* would break up into several parameters.

The last Factor *J* is characterized by concern or guilt over masturbation, conflict between sexual impulses and beliefs, concern over homosexual tendencies, and impulses to perform irrational unnecessary acts which are resisted only with considerable discomfort. It seems clear that *J* is the same as the corresponding factor earlier identified. It represents a conflict within the patient between his sexual impulses and his moral and social standards. The conflict finds expression periodically in overt repetitive acts.

The Second-Order Factors

The correlations between the ten primary vectors were computed, rearranged into two clusters, and factored by the multiple group procedure to yield two orthogonal factors not requiring rotation (2). To facilitate interpretation of the second-order factors, the correlations of the fifty scale variables with these two factors were computed by means of a suitable matrix transformation (5).

Table 3 presents the scale correlations with Factors *AA* and *BB*. Factor *AA* is defined by an anxious tension, a lowered sense of personal adequacy, much preoccupation with insistent thoughts and morbid fears, and to a lesser extent by conflict over sexual problems. The major element seems to comprise an agitated, intropunitive self-preoccupation. The pattern suggests a type of personality disorganization and conflict that occurs when ego defenses crumble.

Factor *BB* is independent of *AA* and represents almost as much of the variance as the latter. *BB* is most strongly defined by primary factors *B*, *E*, and *F*. Hostility, resentment of authority, suspicion and defensiveness are joined with a strong drive for long-term goals and a conforming conscientiousness. Compared to *AA*, this pattern represents a tight defensive system with extropunitive elements.

Table 3

Correlations of the 50 Scales on the Second-Order Factors (decimal points are omitted)

Scale No.	Factors		Scale No.	Factors	
	AA	BB		AA	BB
1	−04	44	26	34	−06
2	14	−09	27	55	02
3	−14	23	28	46	−17
4	−11	12	29	38	10
5	15	29	30	−06	32
6	10	30	31	−31	30
7	−36	33	32	−13	06
8	−05	60	33	03	35
9	22	37	34	18	36
10	−07	49	35	03	17
11	−15	53	36	00	51
12	03	55	37	20	45
13	−21	22	38	13	−15
14	47	02	39	11	39
15	52	02	40	−02	−11
16	28	04	41	24	−06
17	49	04	42	00	−14
18	29	18	43	06	00
19	37	−07	44	14	05
20	28	−10	45	13	−19
21	09	09	46	−04	−09
22	19	41	47	37	09
23	50	07	48	28	02
24	14	−21	49	24	18
25	35	−24	50	36	−07

Discussion

In terms of our initial aim to find a set of measures more descriptive of the principal dimensions of personality and mental illness than currently available, there is one pertinent question that might be asked. To what extent do these factors fit into currently accepted personality theory? It would seem easy, perhaps deceptively too easy, to reconcile these factors, in terms of both origin and dynamics, with a number of different theoretical frameworks. The difficulty is not really in linking these factors, as parameters of personality functioning, with one or another of the current theories but rather in choosing from among an embarrassing number of equally plausible constructions. This is perhaps the same problem the clinician has when confronted with his test data and before he formulates his explanation of the dynamics of the patient. Alternate hypotheses are readily available and there is little doubt, especially if he talks about "basic" dynamics, his ultimate choice depends as much, if not more, on his own theoretical persuasion as on the goodness of fit of data to hypothesis. However, in the present study, primary interest was focused on establishing the factors more firmly on an independent second sample. Rather than speculate with little supporting data it seemed preferable to leave for another study the experimental testing of one or more hypotheses that might be proposed to explain the parameters identified.

The problem of theoretical formulation is further complicated by the fact that the factors cannot be considered in isolation. It seems highly likely that they operate configurally in the individual. Thus, a high level of motivation for long-term goals (Factor *E*) is commonly regarded as a positive trait. But it is not so clearly positive if it is associated with restraint of emotional expression (Factor *A*), a hostile rebelliousness (Factor *B*), and an excessive conscientiousness (Factor *F*). Similar complex interactions can be noted for most of the other factors. The organization of these parameters in the various personality types must be considered, although each factor is relatively independent in the population as a whole.

Another problem that merits discussion concerns the relation of the factors identified in this population to those likely to be obtained for a "normal" group of adults. It is our presumption that most neurotic behavior is on the same continuum but is an extreme in either direction from "normal" behavior. For example, an excess of restraint in expression of feeling and emotion and the lack of appropriate control of emotional expression represent behavior deviations. Both are extremes of a trait of emotional responsiveness. If this assumption is valid, and there is much to support it, then the ten personality parameters represented by our factors are also descriptive of normal individuals.

Summary

This study sought to confirm and clarify ten personality and symptom patterns previously isolated. A sample of 215 male World War II and Korean War veterans were rated by their therapists on 61 scales descriptive of patient behavior, symptoms, and certain inferred needs and attitudes. Seven of the factors hypothesized were fully confirmed while three factors were regarded as related to the original factors only in part. An analysis of the correlations between the primary factors resulted in the identification of two interpretable second-order factors.

Received November 1, 1955.

References

1. American Psychiatric Association Committee on Nomenclature and Statistics. *Diagnostic and statistical manual of mental disorders.* Washington, D. C.: Amer. Psychiat. Ass. Mental Hospital Service, 1952.
2. Harmon, H. H. The square root method and multiple group methods of factor analysis. *Psychometrika,* 1954, **19**, 39–55.
3. Lorr, M., & Rubinstein, E. A. Factors descriptive of psychiatric patients. *J. abnorm. soc. Psychol.,* 1955, **51**, 514–522.
4. Lorr, M., Schaefer, E., Rubinstein, E. A., & Jenkins, R. L. An analysis of an outpatient rating scale. *J. clin. Psychol.,* 1953, **9**, 296–299.
5. Thurstone, L. L. *Multiple-factor analysis.* Chicago: Univer. of Chicago Press, 1947.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

Social Values and Psychotherapy[1]

Richard Jessor
University of Colorado

While problems of professional ethics and of technique in psychotherapy have received much attention from clinical psychologists, almost no consideration has been given to *value* problems involved in carrying on treatment. The main points to be developed in this paper are that value judgments or positions are inescapable in psychotherapeutic practice; that such judgments to date have been largely implicit and unrecognized; that they are based upon a relatively narrow conception of adjusted or healthy behavior; and that a major goal of psychotherapy should involve *social* change.

Perhaps the first value decision which the psychologist faces is whether to do therapy or not. It has been argued by some that it is improper to attempt to change other people's behavior—even though they request assistance—that it is unethical to interfere with the "natural" course of events. Stated this boldly, the problem hardly seems to have created conflicts for psychotherapists. The obvious solution to this problem, as stated, lies in recognizing that the position of *non*-interference in other persons' behavior is no less of a value judgment and requires no less justification than the decision to interfere, i.e., to do therapy. Furthermore, the value position of interference is clearly congruent with broader values in our culture, such as those involved in the process of education.

Some psychologists, incidentally, have not been comfortable with the foregoing position. They have, instead, attempted to avoid the problem by adopting a therapeutic role which minimizes the therapist's interference. Such a position does not seem to escape responsibility for intervention. It simply means that the therapist's role may vary widely from a very active to a very passive one. Nonetheless, at every point on this dimension, the same basic value position must have been taken, namely, that some change in the patient is to be facilitated, i.e., that therapeutic interference of some sort is necessary or desirable. Once this is done, the question of how best to achieve such behavior change—whether by active or passive therapy—becomes an empirical one.

A more difficult problem involving social values is the question of what *direction* therapy should take. In medicine the norm of health is a biological one and relatively clear-cut. The physician is rarely troubled about the direction treatment for physical illness should take. In psychology, however, the norm of health or adjustment is to a large extent relative and variable. It is generally undefined or else defined only implicitly within various theoretical approaches to personality. The therapist, in every case however, makes a choice of therapeutic direction, either explicitly or, most often, implicitly by working within the framework of a particular theory. It is the nature of these choices which constitutes a serious value problem in psychotherapy.

Discussions of psychotherapy in nearly all recent texts refer to two general directions or goals of treatment. The first of these is *intrapersonal* and usually refers to the reduction or elimination of conflict or inner tension, or some such subjectively-oriented concept. Little or no concern is shown for the social origins, implications, or context of behavior. Perhaps the best example of this view is expressed by Carl Rogers (9). He insists that the therapist must be willing to give the client full freedom as to outcomes, including the choice of antisocial or immoral goals, regression, and even death. While Rogers assumes that such complete freedom will inevitably lead in a positive direction, it is apparent that such an extremely intra-individual position as to the direction of therapy overlooks the fact that adjustment is in part socially defined and certain choices are simply not open to the client.

---

[1] An earlier version of this paper was read at the 1955 meeting of the APA in San Francisco.

This position is usually tempered by reference to the second major goal of treatment—what Cameron (2) has called the agreement of behavior with cultural expectations; what others have referred to as *social conformity*. There are many difficulties in accepting this goal, too. Society is not an undifferentiated structure, neither is it static. The question must be asked, to what groups of society is the individual to conform, to which aspects of a *changing* society? In an earlier publication, Rogers recognized these problems. He wrote: "There is no doubt that in some instances social work and even clinical work has been used as a means of blocking social progress. Unless the worker with individuals is alive to the significant movements and trends in our present-day culture, he may easily be seduced into upholding some fixed notion of the socio-economic situation" (8, p. 355).

Most therapists are middle-class individuals. The studies of Robert Havighurst (6), and August Hollingshead (7) offer clear-cut evidence of the influence of the values and social goals of middle-class teachers (their emphasis on neatness, manners, and politeness) on their teaching practices and relations with students. It is not unlikely that the same is true in psychotherapy. Some interesting related evidence on this point has recently been published. At one outpatient clinic a significant relationship was found between the patient's social class and the decision of the intake conference. In a study by Schaffer and Myers (10) it was found that whereas 64.7 per cent of patients from the professional and executive class were assigned for therapy to senior psychiatric staff and residents, only 2.4 per cent of the lowest class and only 33.4 per cent of the next lowest class received such high-level therapists. A patient of bottom-class status has between five and seven times the likelihood of being not recommended for therapy as does the patient from one of the top classes. The authors of this study mention, among other reasons for their findings, implicit evaluations by therapists as to what sort of patient is really more worth while. Such evaluations need to be made explicit in order to determine whether they reflect mere ethnocentric bias or some broad social orientation.

Furst (5) has stressed the implicit values regarding social conformity contained in psychoanalytic theory. He cites, for example, the interpretation frequently given neurotic women that their difficulties with men stem from "penis envy," and suggests that such interpretations tend to exonerate society for the historically inferior position of women.

Consideration of the social-value implications and shortcomings of the two generally accepted goals suggests the need to include a third major goal in psychotherapy—what, for want of a better term, may be called *social contributiveness*. This goal, too, represents a particular value orientation and is not without difficulties, e.g., how to get agreement on what constitutes socially contributive behavior. Nevertheless, its components tend to balance the subjective and static orientations of the other two concerning adjustment.

Two aspects of social contributiveness may be mentioned. One involves the notion expressed by Cantril (3) and L. K. Frank (4) that much personal conflict is due to objective conflicts, contradictions, and shortcomings in the socioeconomic organization in which the patient lives. Therapy then ought to influence the patient to work toward the elimination of objective social conflicts or toward cultural reorganization in order to escape his own personal conflicts. The other aspect of social contributiveness stresses the possible importance, for stability of personality, of a broad social philosophy. This essentially is Adler's (1) emphasis on the importance in therapy of developing what he called social feeling or cooperative work for the benefit of others. More recently Weisskopf-Joelson (11) has made a similar point. She feels a consistent *Weltanschauung* to be important for adjustment and notes its absence in modern Western culture. The contemporary stress on flexibility and tolerance of ambiguity as symptoms of good adjustment suggests to her "a subterfuge of an unstable culture which has little to give but ambiguity" (11, p. 604).

To summarize, it is suggested that psychotherapists direct part of their efforts toward increasing the social contributiveness of patients in two ways:—(*a*) by helping them develop broader social feelings, and (*b*) by encouraging active participation in the process of constructively changing their social environments. Teaching the nature and importance of social contributiveness makes therapy coterminous with education as a whole.

The burden of this position, of course, rests upon the determination of what is socially contributive behavior, a difficult but not insuperable problem. In many cases, e.g., where social discrimination or chauvinism is involved in the disturbance, the answer is obviously contained in explicit values which are part of the general democratic tradition. For the future, however, this viewpoint would suggest two things. One is that consideration of the whole problem of social values involved in psychotherapy be part of the *training* which clinical psychologists (and, of course, psychiatrists) receive. The second is that some multidisciplinary group, e.g., SPSSI, might well undertake a broad study of the issue of socially contributive behavior. The study of values need not be impervious to empirical research. For one thing, if we can get agreement on distant or more general values or goals, then the relationship of more immediate or specific values to those distant ones should be empirically determinable.

Received October 24, 1955.

References

1. Adler, A. *Social interest: a challenge to mankind*. London: Faber & Faber, 1938.
2. Cameron, N., & Magaret, Ann. *Behavior pathology*. Boston: Houghton Mifflin, 1951.
3. Cantril, H. The place of personality in social psychology. *J. Psychol.*, 1947, 24, 19–56.
4. Frank, L. K. *Society as the patient*. New Brunswick: Rutgers Univer. Press, 1948.
5. Furst, J. B. *The neurotic: his inner and outer worlds*. New York: Citadel Press, 1954.
6. Havighurst, R. J., & Taba, Hilda. *Adolescent character and personality*. New York: Wiley, 1949.
7. Hollingshead, A. de B. *Elmtown's youth*. New York: Wiley, 1949.
8. Rogers, C. R. *The clinical treatment of the problem child*. Boston: Houghton Mifflin, 1939.
9. Rogers, C. R. *Client-centered therapy*. Boston: Houghton Mifflin, 1951.
10. Schaffer, L., & Myers, J. K. Psychotherapy and social stratification. *Psychiatry*, 1954, 17, 83–93.
11. Weisskopf-Joelson, Edith. Some suggestions concerning Weltanschauung and psychotherapy. *J. abnorm. soc. Psychol.*, 1953, 48, 601–604.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

Stability of Interaction Patterns During Interviews: A Replication[1]

Joseph D. Matarazzo, George Saslow,
Massachusetts General Hospital and Harvard Medical School[2]

and Samuel B. Guze
Washington University School of Medicine

In an earlier study utilizing 20 subjects (5) we reported that interaction patterns for any given individual during a partially standardized psychiatric interview had markedly high stability coefficients across two interviewers. In addition to demonstrating this stability of patient interaction patterns (effects) when the stimulus (interviewer's behavior) was controlled along certain dimensions, it was shown that these same highly stable interviewee patterns could be reliably modified by planned changes in the interviewer's behavior from one subperiod of the interview to another.

The method of measurement used was the interaction chronograph, an instrument devised by Chapple (1, 2) for recording certain temporal aspects of the interactions between an interviewee and interviewer. A more complete description of the variables studied and the method of observation used will be found in our previous report (5), while a detailed review of the literature, a history of the development of the interaction chronograph, and the theory underlying it will be found elsewhere (4).

Briefly, the standardized interview involves certain "rules" for the interviewer to follow during each of five predefined subperiods of the interview; Periods 1, 3, and 5, consisting of *free* give-and-take interviewing, and Periods 2 (silence) and 4 (interruption) involving *stress* phases of the interview. Through the interviewer's varied behavior during these five periods, the standardized interview presents to an interviewee a sequence of different behaviorial situations designed to elicit his characteristic interaction patterns in what can be thought of as varying miniature interpersonal situations. It should be noted that the stress of Periods 2 and 4 is provided by the overt behavior ("not responding" and "interrupting," respectively) of the interviewer, and not by the "content" of what he says to the interviewee. The durations of Periods 1, 3, and 5 are *fixed* (10, 5, and 5 minutes, respectively); whereas the durations of Periods 2 and 4 are *variable* (12 failures to respond or 15 minutes, whichever is shorter, and 12 interruptions or 15 minutes, whichever is shorter, respectively).

Detailed definitions of the interaction variables investigated in this and the previous study can be found in our earlier reports (4, 5). Briefly they are: (*a*) *A's Units*, the *frequency* of the patient's actions; (*b*) *Action*, the average *duration* of A's actions; (*c*) *Silence*, the average *duration* of A's silences; (*d*) *Tempo*, the average *duration* of each action *plus* its following inaction as a single measure; (*e*) *Activity*, the average *duration* of each action *minus* its following inaction as a single measure; (*f*) *A's Adjustment*, the average *duration* during which A "interrupted" B *minus* the duration during which A "failed to respond to" B; (*g*) *B's Adjustment*, the average *duration* of the other (usually the interviewer) person's adjustment; (*h*) *Initiative*, the *relative frequency* with which A initiated to B following a double silence, as in Period 2; (*i*) *Dominance*, the *relative frequency* with which A dominates B in a double action (interruption), as in Period 4; (*j*) *A's Synchronization*, the *frequency* with which A failed to synchronize with B either by failing to respond to B or by interrupting B; (*k*) *B's Synchronization,* the *frequency* of the other (usually the interviewer) person's failure to synchronize with A; and (*l*) *B's Units,* the *frequency* of the interviewer's actions.

---

[1] This investigation was supported by a research grant (M-1107) from the National Institute of Mental Health, of the National Institutes of Health, U. S. Public Health Service.

[2] The data of this study were collected while all three authors were at the Washington University School of Medicine.

Since the reliabilities (stability coefficients) obtained for these variables in the previous study were unusually high for complex psychological variables such as one finds in the interview, it was decided to attempt to cross-validate the findings. The present study is an exact replication of the earlier one (5). Twenty additional outpatient *S*s were interviewed twice, independently and in counterbalanced order by the same two interviewers who participated in the previous study (one a young internist, the other an older psychiatrist). All features of the experiment were identical: the rooms and the observer were the same; the *S*s were all white; they ranged in age from 18 to 62 (median age of 33); and they had presented complaints typical of the first series of 20 patients. Whereas in the first study there were 11 men and 9 women, the present sex distribution was 8 men and 12 women.

Results

As discussed above, 20 minutes of the standardized interview are fixed and, depending upon the subject's behavior in Periods 2 and 4, up to 30 more minutes, totaling 50 in all, are possible to complete the standardized interview. Table 1 contains an analysis of the actual mean length of the 20 pairs of interviews. The data from the original series are included for purposes of comparison. The results of the cross-validation study are nearly identical with those of the original; again showing that despite individual patient variation (shown equally with the two doctors), the interview lasts, on the average, approximately 32 minutes. This is in spite of the fact that in the original series one patient interview lasted 50 minutes, while in the present series the longest interview lasted only 41 minutes. The results shown in Table 1 also duplicate the finding of no *doctor* effect per se; and show again that there is no evidence of an *order* effect since the mean lengths of all first and all second interviews are identical. Interestingly, the earlier finding of a significant (p of .05) order effect for the *Silence* variable with the two doctors was not cross validated. The value of the covariance F was now found to be an insignificant .88 indicating that both doctors were eliciting comparable mean silence values (as well as values for the other variables) from the 20 patients across the total interview. Also, as in the original study, on replication the *mean* lengths of Period 2 (7.63 and 8.28 minutes) and Period 4 (2.47 and 2.35 minutes) were almost identical for the 20 patients with the two interviewers. The value of rho (.752) for length of Period 2 for each of the 20 patients with the two doctors was significant at the .01 level, whereas the extremely limited range for Period 4 made the obtained value of rho (.008) of little practical use.

Table 2 presents, for the 20 patients, the means, standard deviations, and ranges (for the total interview) of nine of the interaction chronograph variables. Values are given for each doctor, with Pearson coefficients of correlation (stability coefficients) shown in the second column from the last. Pearson coefficients of correlation, shown in parentheses, are reproduced from the original study for

Table 1

Analysis of Mean Length of Total Interview in Minutes for Two Series of 20 Patients Interviewed Twice

Analysis	Original Series		Replication Series	
	Mean	Range	Mean	Range
Length of all 40 interviews	32.8	25.7 to 50.3	31.5	24.5 to 41.1
a. Length of all first interviews	32.9	26.9 to 41.4	31.3	24.5 to 40.4
b. Length of all second interviews	32.8	25.7 to 50.3	31.6	25.6 to 41.1
c. Length of all first doctor's interviews	33.5	26.9 to 50.3	30.9	24.5 to 40.4
d. Length of all second doctor's interviews	32.2	25.7 to 41.2	32.0	25.6 to 41.1

Table 2

Means, Standard Deviations, Ranges, and Coefficients of Correlation Across Total Interview for Major I-C Variables

Variable	Dr. 1	Dr. 2	*r*	*p*
1. Mean Pt.'s Units	68.30	76.65	.926	.01
SD	25.95	26.69	(.747)†	
Range	39 to 132	41 to 133		
2. Mean Pt.'s Action*	54.65	41.75	.786	.01
SD	36.06	23.83	(.956)	
Range	12 to 130	11 to 109		
3. Mean Pt.'s Silence	9.55	9.10	.744	.01
SD	3.16	2.26	(.759)	
Range	5 to 19	6 to 15		
4. Mean Pt.'s Tempo	53.65	48.55	.905	.01
SD	22.33	19.61	(.908)	
Range	19 to 100	19 to 100		
5. Mean Pt.'s Activity	35.00	30.35	.891	.01
SD	23.35	20.28	(.930)	
Range	4 to 87	3 to 85		
6. Mean Pt.'s Adjustment	−1.08	−.66	.853	.01
SD	1.43	1.29	(.802)	
Range	+.50 to −5.64	+.88 to −4.95		
7. Mean Dr.'s Adjustment	−2.24	−1.93	.749	.01
SD	1.21	.69	(.737)	
Range	−.77 to −4.82	−.88 to −2.97		
8. Mean Pt.'s Synchron.	.84	.85	.741	01
SD	.08	.05	(.726)	
Range	.69 to .97	.77 to .98		
9. Mean Dr.'s Units	58.75	63.45	.909	.01
SD	26.16	24.94	(.772)	
Range	28 to 119	32 to 115		

* Values for variables 2 through 7 are given in hundredths of a minute. To convert to seconds, multiply the given value by 0.6.

† Values in parentheses are reproduced from original study for comparison.

comparison. The values of *r* for the replication series, ranging from .741 to .926, again show a striking *stability* in interaction variables from first to second interview for each of the 20 patients. Furthermore, as can be seen from the means, standard deviations, and ranges, this stability is *absolute* as well as *relative*. Thus, for example, the ranges for "Pt.'s Units" show that during the same approximately 32-minute standardized interview one *S* interacted 39 times with Doctor 1 and 41 times with Doctor 2, whereas a second *S* interacted 132 times with Doctor 1 and 133 times with Doctor 2. Similar comparisons can be made for each of the other variables shown in Table 2.

In the original study, "Action" was the most stable variable with an *r* of .956: on replication this value dropped slightly to .786. On the other hand, "Pt.'s Units," with an *r* of .926, was now the most stable variable. Despite these slight (statistically nonsignificant) differences, however, all reliability coefficients on replication again were found to be significant at the .01 level of confidence. Thus using predefined and partially standardized interview conditions, and taking the *total* interview as the sample, patients show markedly similar interaction patterns from one interviewer to the other, i.e., independent of which doctor is conducting the interview.

Table 3 demonstrates that interaction patterns are invariant only as long as stimulus (interview) conditions remain the same across the two interviewers for the total interview. As any single interviewer varies his own behavior *within* the interview—from one subperiod to another—there is a corresponding change in the means obtained by patients with him for any one variable. Thus, for example, the *average* interviewee values for "Silence" with Doctor 1 for Periods 1 to 5, respectively, vary from 9.15 up to 15.05 and back down to 9.40 and down still further to 5.65 and finally up again to 8.50 hundredths of a minute as Doctor 1's own silence behavior varies (see "rules" in earlier report)

Table 3

Means and Standard Deviations for Two Interviewers for Each of the Major I-C Variables for Each Period in Replication Series

Period		Pt.'s Units		Pt.'s Action		Pt.'s Silence		Pt.'s Tempo	
		Dr. 1	Dr. 2	Dr. 1	Dr. 2	Dr. 1	Dr. 2	Dr. 1	Dr. 2
I	M	19.50	22.65	68.40	54.60	9.15	9.05	77.90	63.95
	SD	12.06	13.11	46.16	31.32	2.22	2.44	47.29	31.64
II	M	12.90	13.30	44.25	48.80	15.05	13.10	59.05	62.75
	SD	.77	.84	29.46	31.97	9.06	4.50	28.05	30.01
III	M	9.95	11.85	87.40	53.05	9.40	9.30	97.10	62.70
	SD	6.79	6.44	87.30	37.52	4.21	2.90	87.19	38.04
IV	M	14.85	14.50	11.25	9.50	5.65	6.20	16.95	16.05
	SD	3.36	2.89	5.71	2.38	1.28	2.25	6.30	2.40
V	M	11.10	14.35	62.50	42.80	8.50	8.05	71.10	51.10
	SD	6.13	6.52	49.35	31.44	1.80	2.22	48.56	31.55
Total	M	68.30	76.65	54.65	41.75	9.55	9.10	53.65	48.55
	SD	25.95	26.69	36.06	23.83	3.16	2.26	22.33	19.61

Period		Pt.'s Activity		Pt.'s Adjust.		Dr.'s Adjust.		Pt.'s Synchro.		Dr.'s Units	
		Dr. 1	Dr. 2	Dr. 1	Dr. 2	Dr. 1	Dr. 2	Dr. 1	Dr. 2	Dr. 1	Dr. 2
I	M	59.55	45.65	−1.16	−.90	−.39	−.38	.99	1.00	18.85	24.40
	SD	45.15	31.12	1.49	1.68	.25	.18	.03	.00	11.56	12.17
II	M	29.75	36.25	−1.24	−.48	−11.15	−10.50	.26	.26	3.40	3.35
	SD	33.29	33.86	2.40	.57	4.96	3.63	.18	.12	2.31	1.62
III	M	78.30	43.85	−1.56	−1.09	−.52	−.44	1.01	1.00	9.90	11.10
	SD	87.70	37.00	3.39	1.78	.26	.14	.06	.03	6.80	5.73
IV	M	5.70	3.30	−.60	−.44	1.38	1.28	.99	1.00	15.55	14.05
	SD	5.48	3.80	1.03	1.75	.70	.56	.03	.04	3.43	2.42
V	M	54.15	34.90	−1.04	−.57	−.33	−.33	1.00	1.00	11.00	13.55
	SD	50.00	31.52	1.38	1.43	.22	.13	.02	.00	6.07	6.00
Total	M	35.00	30.35	−1.08	−.66	−2.24	−1.93	.84	.85	58.75	63.45
	SD	23.35	20.28	1.43	1.29	1.21	.69	.08	.05	26.16	24.94

Table 4

Correlations for Two Interviewers for Each of the Major I-C Variables for Each Period in Replication Series

Period	Pt.'s Units		Pt.'s Action		Pt.'s Silence		Pt.'s Tempo		Pt.'s Activity	
	r	rho	*r*	rho	*r*	rho	*r*	rho	*r*	rho
I	.909	.870	.711	.870	.618	.574	.706	.870	.718	.792
II	.201	.474	.828	.765	.613	.732	.857	.746	.776	.651
III	.788	.697	.428	.655	.481	.641	.407	.650	.455	.630
IV	.578	.430	.676	.617	.233	.179	.215	.270	.728	.784
V	.823	.817	.662	.821	.555	.484	.666	.824	.658	.776
Total	.926	.917	.786	.914	.744	.828	.905	.941	.891	.910

Period	Pt.'s Adjust.		Dr.'s Adjust.		Pt.'s Synchron.		Dr.'s Units	
	r	rho	*r*	rho	*r*	rho	*r*	rho
I	.713	.733	.254	.525	.000	.850	.909	.842
II	.051	.375	.768	.808	.537	.591	.537	.594
III	.785	.565	−.129	−.023	−.027	.671	.783	.700
IV	.307	.227	.576	.591	.161	.546	.503	.395
V	.470	.306	.510	.614	.000	.793	.807	.830
Total	.853	.821	.749	.780	.741	.717	.909	.880

Note.—For $r = .444$, $p < .05$; for $r = .561$, $p < .01$. For rho $= .427$, $p < .05$; for rho $= .543$, $p < .01$.

from responding quickly (within one second of the patient's last utterance) in Period 1 to not responding for 15 seconds in Period 2, to responding within one second in Period 3, to interrupting the patient in Period 4, and to responding within one second again in Period 5. A similar pattern is evident for Doctor 2. The value for *F* across these 5 correlated means (3, pp. 284–296) was 14.80 for Doctor 1 and 23.44 for Doctor 2, both significant at the .001 level of confidence. Examination reveals that "Action" as a patient interaction variable is also a function of what the interviewer is doing during a particular phase of the interview. Thus when the interviewer (Doctor 1, for example) was conducting a free give-and-take interview in Periods 1, 3, and 5, the patients' *average* duration of utterance ("Action") was 68.40, 87.40, and 62.50, whereas while he was interrupting them in Period 4, the patients' average action fell dramatically to a value of 11.25 hundredths of a minute. Interviewer's silence in Period 2 likewise was associated with a slight drop in patient action to an average value of 44.25. A similar statistically significant (p of .001 in both cases) pattern is evident for patients' "Action" with Doctor 2. It should be pointed out that we are here dealing with *averages*,[3] since Chapple divides all patient variables in a given period by the number of patient units in that period, thus

[3] Analysis of the *individual* records of the 20 patients reveals that, independent of each patient's own base line, all 20 showed the drop in "Action" from Period 3 to Period 4 with a subsequent increase in Period 5. This pattern was true for all 20 patients with Doctor 1 as well as Doctor 2. The picture is less striking for increased "Silence" in Period 2 relative to Periods 1 and 3; with Doctor 1, only 15 out of the 20 showed the group pattern shown in Table 3. Interestingly, the same 5 patients (and only these) again showed this deviation from the group trend with Doctor 2, thereby indicating the "stability" of this pattern despite this deviation from the group average.

controlling for the differences in the lengths of the five periods, as well as controlling for individual differences in interaction rates among patients.

The values shown in Table 3 are thus a replication of the earlier finding that means of interaction patterns of patients, while invariant under identical interinterviewer conditions across the interview as a whole (Table 2), are markedly susceptible to changes in the stimulus conditions (presented as intra-interview changes from Periods 1 to 5 in the behavior of a single interviewer). Thus Table 3 shows that as any single interviewer changes his behavior from one period to another there are associated changes in interviewee interaction patterns; whereas Table 2 shows that a second interviewer can elicit almost identical interviewee patterns (invariance) during this *total* interview providing he changes his behavior in each of the five periods in a manner similar to the first interviewer (i.e., providing both interviewers present identical stimulus conditions to the patient during the 32-minute, on the average, total sample).

The extent to which the mean patient interaction values shown in Table 3 for a single doctor are reliable *for any one period*, in contrast to the interview as a whole, can be seen in Table 4. In the original study, 37 of the 45 period-variable correlations (9 variables times 5 periods) were statistically significant; 27 at the .01 level and 10 at the .05 level. Most of the remaining 8 represented artifacts in measurement; due most frequently to extremely narrow ranges or limited frequencies, and occasionally to skewing. On replication, as seen in Table 4, 32 of the 45 period-variable combinations have statistically significant Pearson r's; 24 at the .01 level and 8 at the .05 level. Values for the rank-order correlation, rho, are also shown for each variable in Table 4. Use of this statistic yields 38 significant values out of the total of 45, with 34 of the 38 reaching the .01 level of confidence. The advantage of rho with these data is shown clearly for the variable "Pt.'s Synchronization" where, due to a limited range, four out of the five period-variables fail to reach significance when r is used and, in contrast, all five reach the .01 level when rho is used. In some cases, then, rho is a better statistic for these data. We have not used it routinely only because, unlike r, it is a terminal statistic and does not require computation of measures such as means and variances which can be used in further analyses. The values of rho for the total interview are given at the bottom of Table 4 and these can be compared with the values of r copied from Table 2 for the same variables. However, whether r or rho is used, it is well to remember that the correlations shown in Table 4 are based on only a small fraction of the total interview sample (e.g., Period 3 correlation values represent only 5 minutes of the total 32-minute interview). One is therefore all the more struck by the similarity or stability of interaction patterns exhibited by the 20 patients with the 2 interviewers.

Finally Table 5 shows the stability coefficients for the "Initiative" (Period 2) and "Dominance" (Period 4) variables in both the original and present studies. These are values for r; values for rho are *.556* and *.713* for Initiative and Dominance, respectively, both significant at the .01 level. Although on cross validation there is a reversal in their relative magnitudes, both variables show considerable stability.[4] When one remembers that Period 2 (silence stress) lasted only 7 to 8 minutes on the average, and Period 4 (interruption stress) only slightly over 2 minutes, these stability coefficients are quite striking. The degree of invariance of a single individual's "initiative" or "ascendance-submission" patterns could be investigated in greater detail by extending in a separate study the lengths of these two periods to perhaps 15 minutes each. It is reasonable to assume that a larger sample of these behaviors would be associated with a higher stability coefficient.

Before leaving Table 5, it should be noted again that, on the average, patients take the initiative more than they have to be initiated

---

[4] Our earlier suggestion that the reliability of Dominance, which was lower than Initiative in the first study, might be a function of the "faster pace" of Period 4 relative to Period 2 is not borne out by this reversal in the relative magnitudes of r for these 2 variables.

Table 5

Means, Standard Deviations, Ranges, and Coefficients of Correlation for Two Interviewers for Initiative (Period II) and Dominance (Period IV) in Original and Replication Series

	Original Series			Replication Series		
Variable	Dr. 1	Dr. 2	r	Dr. 1	Dr. 2	r
1. Mean Initiative	.75	.77	.805**	.72	.73	.552*
SD	.18	.19		.18	.14	
Range	.17 to .94	.23 to 1.00		.25 to .92	.42 to .92	
2. Mean Dominance	−.32	−.42	.470*	−.36	−.43	.697**
SD	.31	.30		.31	.26	
Range	−.71 to +.83	−.92 to +.29		−1.00 to +.13	−.92 to +.00	

* Significant at the .05 level of probability.
** Significant at the .01 level of probability.

to (means of + .72 and + .73) under conditions of silence, and submit more than they dominate (− .36 and − .43) under conditions of interruption. The fairly large standard deviations for each of these variables indicate that these means are only group trends and that individuals vary considerably in these two dimensions of their interaction patterns.

Discussion

The results of this cross validation would indicate that interaction patterns such as the ones we have studied, as one dimension of personality, are both modifiable by planned changes in intra-interviewer behavior and remarkably stable from one interviewer to another in interviews conducted on the same day. This finding may be a function of many factors, one of which may be the "set" of the patient on the particular day he has his two interviews. If this is true, one would expect the patient's interaction pattern possibly to be different if the two interviews were separated in time. We have completed a study, again on 20 subjects, in which a *single* interviewer interviewed the same patient on two occasions one week apart. The results of this study, now being analyzed, should provide us with information which will enable us to decide the usefulness of the interaction chronograph as an instrument for measuring changes of behavior *in time* (for example, as a function of interpolated psychotherapy, etc.). However, the results of the two studies already completed seem to indicate that the interaction chronograph has promise as an instrument for assessing changes in behavior during a single day (test-retest changes in behavior as a function of the administration of a drug, for example).

Another factor which is of theoretical as well as practical importance is the effect of the *content* of the interview on the interaction patterns. We have been working on a method of content analysis and hope soon to be able to compare, for these same interviews, *what* the interviewee said with *how* he said it, i.e., "content" vs. interaction chronograph pattern. In this manner we plan to investigate Chapple's hypothesis that interaction patterns are relatively independent of content.

Summary

The present investigation, a replication of an earlier one, was a study of the interaction patterns of 20 patients who were interviewed independently by two interviewers on the same day. The demonstrated cross validation of our earlier findings indicates that a single patient's interaction pattern is both susceptible to planned changes in a single interviewer's behavior (*intra-interviewer* modifiability) and that this pattern is remarkably stable, or invariant, from one interviewer to another (*interinterviewer* stability). Research currently in progress is designed to investigate

further aspects of this time dimension of interaction patterns.

Received February 21, 1955.
Early Publication.

References

1. Chapple, E. D. The interaction chronograph; its evolution and present application. *Personnel,* 1949, **25**, 295–307.
2. Chapple, E. D. The standard experimental (stress) interview as used in interaction chronograph investigations. *Hum. Organization,* 1953, **12** (2), 23–32.
3. McNemar, Q. *Psychological statistics.* New York: Wiley, 1955.
4. Matarazzo, J. D., Saslow, G., & Matarazzo, Ruth G. The interaction chronograph as an instrument for objective measurement of interaction patterns during interviews. *J. Psychol.,* 1956, **41**, 347–367.
5. Saslow, G., Matarazzo, J. D., & Guze, S. B. The stability of interaction chronograph patterns in psychiatric interviews. *J. consult. Psychol.,* 1955, **19**, 417–430.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

The Relation of Brain Injury and Visual Perception to Block Design Rotation

Harold L. Williams, Ardie Lubin, Charles Gieseking
and Irvin Rubinstein
Walter Reed Army Institute of Research

This report concerns the phenomenon called rotation which sometimes occurs on block-design tests or other visual-motor tasks where designs have to be reproduced. The purpose of the investigation was to repeat and extend the studies of rotation reported by Shapiro (1, 2, 3) and Yates (4). In a recent series of papers they discussed the effects on rotation of such factors as changes in geometric properties of the target, blocking the subject's (*S*'s) peripheral vision, intelligence of the *S*, and brain injury.

In the Block Design Rotation Test (BDRT), which was devised by Shapiro, the *S*'s task is to reproduce a blue and yellow design, one inch square, painted on a white card six inches square. Each design requires four blocks taken from the Wechsler Bellevue Block-Design subtest. Rotation (*R*) is measured as the number of degrees which the *S*'s reproduced design differs in orientation from that of the target design. These scores are summed over all forty cards.

The BDRT was constructed especially to test the assumed relation between rotation and certain geometric properties of the target designs. These properties arise from the nature of the ground, the figure, and the type of symmetry involved. The ground—the white card on which the design is painted—is presented either as a square or a diamond. The figure, or design, is also presented either as a square or diamond. The line of symmetry is an imaginary line which cuts the design into mirror images. Two kinds of symmetry are possible: rectanglar, in which the line of symmetry is vertical or horizontal, and diagonal, in which the line of symmetry is 45 degrees from vertical. In Figure 1*a*, for example, the target has a square ground, square figure, and diagonal symmetry. In Fig. 1*b*, it has square ground, diamond figure, and rectangular symmetry.

Over the 40 cards with their 10 basic designs, all possible combinations of symmetry, figure, and ground occur an equal number of times. They are presented in a balanced order developed by Joan May (2, p. 617).

Shapiro's findings concerning the geometric properties of the target can be summarized as follows: (*a*) diagonal symmetry produces more rotation than rectangular symmetry; (*b*) diamond figure produces more rotation than square figure; (*c*) diamond ground produces more rotation than square ground; (*d*) the order of effect of the three properties is symmetry, figure, and ground; and (*e*) the effects of symmetry, figure, and ground are additive.

Shapiro, and later Yates, found that brain-injured *S*s rotated significantly more than control *S*s. About 75 per cent correct classifi-

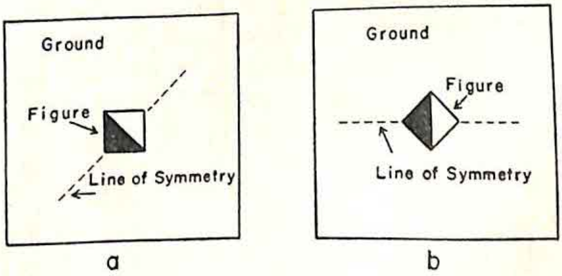


Fig. 1. Two possible combinations of symmetry, figure, and ground.

cation can be obtained with their data when *R* is used as a predictor variable.

With the finding that the factors of symmetry, figure orientation and ground orientation appear to have directional properties which affect the amount of rotation, Shapiro developed a theory to account for the relationship between brain injury and rotation. He argued that if one could assume that the brain-injured *S*s tended to focus on the central part of the target at the expense of the periphery, then directional properties of the target would presumably exert a more powerful influence toward rotation than the directional properties of the more peripheral background. He then deduced that non-brain-injured *S*s, deprived of peripheral vision, would rotate as much as brain-injured *S*s, and that they would respond to the same geometric properties of the target (symmetry, figure, and ground). To test this deduction he fashioned a mask (field reducer) which covered one eye but permitted vision with the other eye through an aperture 6 mm. in diameter. Control *S*s wearing the field reducer and working on a plain black felt surface rotated as much as brain-injured *S*s and responded to the same geometric properties of the target, thus confirming the hypothesis.

Three main problems are considered here: (*a*) the effect of type of symmetry, orientation of figure, and orientation of ground on rotation; (*b*) the effect of blocking peripheral vision when Shapiro's experiment is extended and brain-injured *S*s as well as controls work with a field reducer; and (*c*) the effect of differences in intelligence (a variable which was studied by Shapiro and Yates but which they do not appear to regard as a major explanatory one).

In the present study a group of brain-injured patients was again compared with a group of non-brain-injured controls, and with a control group wearing a field reducer. Another group of brain-injured *S*s wearing the field reducer was added in order to assess the effects of the field reducer in conjunction with brain injury. A group of control *S*s with low intelligence was added in order to explore more thoroughly the effects of intelligence on rotation. Thus the following five groups of 20 *S*s each are involved:

1. Brain-injured without field reducer (BG).
2. Brain-injured with field reducer (BF).
3. Controls without field reducer (CG).
4. Controls with field reducer (CF).
5. Low-intelligence controls without field reducer (CD).

Subjects and Procedure

Selection of Subjects

Forty male brain-injured patients were selected from the Neurology and Neurosurgery Services at Walter Reed Army Hospital. The brain-injured sample was heterogeneous with respect to diagnosis. There were 18 head-injury cases, 13 patients with neoplasms or vascular anomalies, and 9 with diffuse brain pathology of various kinds. The *S*s can be described as young males, in good health prior to illness, who were examined early in recuperation from injury or surgery. In general their brain injuries were acute rather than chronic, and diffuse rather than focal.[1]

Sixty male controls were obtained from among non-brain-injured medical patients. They were heterogeneous with respect to diagnosis, but were ambulatory, and not acutely ill. They had exhibited no neurological symptoms on a routine examination at admission. Twenty of these controls were selected on the basis of low intelligence. Their scores on Aptitude Area I of the Army Classification Battery were at or below the thirty-second percentile.[2]

Not all *S*s were able to complete all designs well enough for rotation to be measured. If *S*s were unable to complete at least three of the five designs in one of the subsets used to test the effects of figure, ground, and symmetry, they were dropped from the study. There were four such *S*s in each of the first four subgroups. Sampling was continued until 20 *S*s per group were obtained. It has not been possible to evaluate the biasing effect of dropping these *S*s.

The field reducer was assigned randomly to 20 of the brain-injured group and to 20 of the controls. Since the effect of eliminating peripheral cues (independent of the working surface) was being tested, all *S*s worked on a black felt cloth spread over the table top.

---

[1] Our thanks are due Lt. Colonel James F. Hammill, Chief, Neurology Service, Walter Reed Army Hospital, who made the final diagnostic decision on each case.

[2] Aptitude Area I is based on scores on verbal type subtests which have been shown by Stone (Stone, H. Personal communication, 15 July 1955) to have substantial correlations with Wechsler Bellevue Full-Scale IQ. The percentile corresponds roughly to an IQ of 80 to 85.

Table 1

Effect of Symmetry, Figure, and Ground on Rotation

	Ground		Figure		Symmetry	
Group	$D > S$*	Significance Level	$D > S$*	Significance Level	$Di > Re$†	Significance Level
BF	14	NS	19	1%	14	NS
BG	12	NS	15	5%	13	NS
CF	10	NS	19	1%	15	5%
CG	9	NS	19	1%	16	1%
CD	15	5%	16	1%	17	1%
Sum	60	5%	88	1%	75	1%

* "$D > S$" refers to the number of *Ss* whose total amount of rotation was greater on the 20 designs with diamond shapes than on the 20 designs with square shapes.

† "$Di > Re$" refers to the number of *Ss* whose total amount of rotation was greater for designs with diagonal symmetry than for designs with rectangular symmetry.

Prior to administration of the BDRT, all *Ss* in the first four groups were given five subtests of the Wechsler Bellevue Adult Intelligence Scale, Form 1. These were Arithmetic, Block Design, Similarities, Vocabulary, and Comprehension. Because of time pressure, the CD group was given only Block Design and Arithmetic.

Procedure

The procedure was essentially the same as Shapiro's. The cards were presented one by one, centered on the table from left to right, and with the edge or corner nearest the *S* placed 12 inches from his edge of the table. The *S* was seated as close to his edge as possible and was told to construct the designs as close to the edge as he could. The experimenter sat opposite the *S*. When the *S* finished a design he was instructed to turn away and close his eyes. At this point, time was recorded and the degrees of absolute rotation measured with a protractor.

Results and Discussion

By way of a summary statement, the following points can be made. The results are in general but not complete agreement with Shapiro's on the relationship between rotation and the various stimulus properties. The effect of blocking peripheral vision in brain-injured *Ss* is not in line with Shapiro's hypothesis, and differences in intelligence appear to be more important than he assumed.

Stimulus Properties and Rotation

To measure the effect of each stimulus property, different subsets of problems were compared with one another. For example, in measuring the effect of ground, an *S*'s rotation score on the 20 cards with a diamond ground was compared with his score on the 20 cards with a square ground. Then the number of *Ss* in each group was determined for whom the first score was higher than the second.

Table 1 summarizes these results and presents also the results of sign tests. The sign test was used rather than the usual *t* test because in all groups, *R* was distributed non-normally with a large positive skew.

Table 1 shows that the diamond figure produces more rotation than the square figure, and diagonal symmetry more than rectangular symmetry. These results are in agreement with Shapiro's. The results on diamond vs. square ground, however, are somewhat equivocal (out of 100 *Ss*, 60 show a difference in the expected direction, but in only the CD group is the difference significant). The combined results are significant, however, at the 5 per cent level.[3]

When Shapiro assessed the relative strength of the geometric properties of the target, he found that, in general, diagonal symmetry was associated with more rotation than diamond figure, and that diamond ground was least effective. Table 1 indicates that, in the present study, figure tended to affect the amount of rotation more than symmetry for four of the five groups. The effect of ground, as Shapiro reported, tended to be weakest.

---

[3] For complete tables of raw data, details of procedure, and description of the sample, as well as for discussion of certain differences between our procedure and Shapiro's, the reader is referred to Walter Reed Army Institute of Research Interim Report (in press) entitled "An Experimental Study of Block Design Rotation."

What about the interaction of symmetry, figure, and ground? Shapiro reasoned that from the Gestalt viewpoint, incongruent combinations should produce more rotation than congruent ones. He took the amount of rotation on trials within congruent figure-ground cards (ten designs with a square figure and square ground, plus ten designs with a diamond figure and diamond ground) and compared it with the amount of rotation on incongruent figure-ground cards (ten designs with square ground and diamond figure, plus ten designs with diamond ground and square figure). If the effects of ground and figure were additive, then there should have been no difference between the two sets. There were no over-all significant differences in his data although there was a slight tendency for incongruence to produce more rotation. Shapiro's analysis was repeated for this sample, and Table 2 shows the results. It can be seen that there is a tendency for incongruent (*I*) figure-ground designs to produce more rotation than the congruent (*C*) figure-ground designs. This tendency appears in each of the groups, and the combined results are significant at the 1 per cent level. The other two interactions do not reach a significant level in any group nor in the combined results.

The present results, then, concerning the relation of rotation to the geometric aspects of the target designs differ from Shapiro's in only two respects: (*a*) figure is found to be more powerful than symmetry, and (*b*) figure-ground incongruence produces significantly more rotation than figure-ground congruence. Thus there appear to be not only independent effects of symmetry, figure, and ground, but also a Gestalt interaction between figure and ground.

The Effect of Blocking Peripheral Vision

In line with Shapiro's results, the present control group showed more rotation with the field reducer than without it ($p < .01$). In disagreement with the extension of Shapiro's hypothesis, rotation by brain-injured with the field reducer is significantly *less* than without the field reducer ($p < .01$).

The scores on which these comparisons are based are given in the first two lines of Table 3, which shows the mean and variance of rotation scores (M) for each of the five groups. (The logarithmic function $M = 100\ [\log R - 1]$ was used in place of R to reduce the very large positive skew.)

It will be recalled that Shapiro suggested that the brain-injured tend to disregard the periphery. If this were so, then blocking the visual periphery should make little difference to the patient, and one would expect about the same amount of rotation in the BF group as in the BG group. The fact that the BF group shows less rotation than the BG suggests that the field reducer may prevent pe-

Table 2

Interaction of Symmetry, Figure and Ground

Group	Figure and Ground		Figure and Symmetry		Symmetry and Ground	
	$I > C$*	Significance Level	$I > C$*	Significance Level	$I > C$*	Significance Level
BF	15	5%	10	NS	8	NS
BG	11	NS	10	NS	12	NS
CF	13	NS	8	NS	12	NS
CG	13	NS	10	NS	11	NS
CD	14	NS	7	NS	8	NS
Sum	66	1%	45	NS	51	NS

* "$I > C$" refers to the number of *S*s whose total amount of rotation was greater on the 20 incongruent designs than it was on the 20 congruent designs.

Table 3

Rotation Scores (M) and Their Relation to Scores (T) on Wechsler Subtests

Statistic	BF	BG	CF	CG	CD	Total
N	20	20	20	20	20	100
$\bar{M}$	131.25	165.55	144.85	115.95	168.80	145.28
$s_M{}^2$	1,271.78	1,114.89	1,352.13	807.21	1,355.54	1,540.97
$\bar{T}$	26.50	23.20	27.25	30.45	20.95	25.67
$s_T{}^2$	111.95	122.90	92.62	69.73	39.63	91.90
r_{MT}	−.56	−.71	−.53	−.38	−.51	−.60
$\bar{M}-M_T$	−12.37	15.34	2.72	−19.79	14.10	

ripheral stimuli from distracting the brain-injured *S*, thus enabling him to concentrate more effectively on the target. In place of Shapiro's "inattention" hypothesis, one would then substitute the notion that the periphery of brain-injured *S*s provides distorted and confusing information about the visual frame of reference.

The Effect of Intelligence

How well can brain-injured *S*s be discriminated from controls? As Table 3 shows, the rotation scores in the BG group are much higher than those in the CG group. (The difference is significant beyond the 1 per cent level.) If the M score of 136 is used as a cutoff, 75 per cent of the *S*s in these two groups can be classified correctly. However, Table 3 also indicates that there is a substantial negative correlation between the M score and a score (T) where T equals the sum of Wechsler Block Design and Arithmetic. Yates also reported significant negative correlations between intelligence and rotation. Since the CG group has higher intelligence scores than the BG group, perhaps not all of the 75 per cent correct classification was due to brain injury per se. This was the reason for adding the dull (CD) group. In fact, when CD is considered part of the control group there is no cutoff M score that will produce usable discrimination between brain-injured and controls. This finding raises the question whether all of the difference between the rotation averages for the five groups may not be due entirely to differences in intelligence. To test this, an analysis of covariance of M was made, holding T constant. That is, each *S*'s M score was statistically equated for the effect of intelligence. The results showed that the rotation means, even when corrected for intelligence, differed significantly beyond the 1 per cent level.

In Table 3 the line labeled $\bar{M} - M_T$ gives the rotation means corrected for the effects of intelligence scores. An immediate contradiction appears. The CD group has a significantly higher average rotation score than the CG group even after correction for intelligence. Since these groups were intended to differ only with respect to intelligence, this suggests that some unknown sampling bias toward higher rotation in the CD group may have existed.

So far, then, it appears that rotation is affected somewhat by the geometric properties of the target card, and a great deal by the intellectual functioning of the *S*. It is difficult to conclude from the data what the direct effect of brain injury is on rotation, since there appears to be an interaction of brain injury with the effect of eliminating peripheral cues.

Summary

The Block Design Rotation Test was given to five groups (with 20 *S*s in each): (*a*) brain-injured patients who wore a field reducer (i.e., eye-mask that restricted peripheral vision); (*b*) brain-injured patients without field reducer; (*c*) non-brain-injured patients without field reducer; (*d*) non-brain-injured patients with field reducer; and (*e*) dull normal patients without field reducer. Most of the relationships given by Shapiro and Yates were confirmed:

1. A diagonal line of symmetry in the target design produced more rotation in the

S's reproduced design than a rectangular (horizontal or vertical) line of symmetry.

2. When the target design was presented as a diamond figure, there was more rotation than when it was presented as a square figure.

3. When the target design was presented on a diamond ground, there was more rotation than when it was presented on a square ground.

4. Both figure and symmetry were more powerful than ground in producing rotation.

5. There was a substantial negative correlation between rotation and various subtests of the Wechsler Bellevue.

The present findings differ from Shapiro's in certain respects:

1. When intelligence scores were held constant, the difference in rotation for brain-injured and controls remained statistically significant but was not great enough to be used for classification.

2. There was a slight but significant Gestalt interaction between figure and ground. Rotation is not only a function of the separate effects of figure, symmetry and ground; it is also increased when figure and ground are incongruent.

3. Reducing peripheral vision increased rotation in controls, and appeared to *decrease* rotation in the brain-injured. This finding suggests that whereas the normal *S* probably uses peripheral cues as guides to correct orientation of his designs, the brain-injured *S* may be confused and distracted by them. Masking his visual periphery assists rather than hinders him in orienting his design.

Received October 10, 1955.

References

1. Shapiro, M. B. Experimental studies of a perceptual anomaly. I. Initial experiments. *J. ment. Sci.,* 1951, **97**, 90–110.
2. Shapiro, M. B. Experimental studies of a perceptual anomaly. II. Confirmatory and explanatory experiments. *J. ment. Sci.,* 1952, **98**, 605–617.
3. Shapiro, M. B. Experimental studies of a perceptual anomaly. III. The testing of an explanatory theory. *J. ment. Sci.,* 1953, **99**, 394–409.
4. Yates, A. J. An experimental study of the block design rotation effect with special reference to brain-damage. Unpublished doctor's dissertation, Univer. of London, 1954.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

A Second Study of Psychological Changes During the First Year Following Prefrontal Lobotomy[1]

John F. Winne and Isidor W. Scherer
VA Hospital, Northampton, Mass.

Scherer and his associates have published two reports (11, 12) dealing with the changes in a sample of schizophrenic patients following prefrontal lobotomy. The present paper reports the results of a similar experiment utilizing an independent sample.

Earlier studies of the effect of prefrontal lobotomy on psychological test scores (1, 2, 4, 7, 9) have yielded conflicting results. Scherer *et al.* (12) suggested that part of this discrepancy might be due to differences in the length of the interval between pre- and posttesting. Other factors, cited from a review of the literature, which might lead to conflicting results included lack of control subjects, heterogeneity of the sample populations under study, variations in the type of operation, and failure to integrate test findings with clinical data. By controlling these factors in an approach which involved successive verification of hypotheses (12, pp. 2–5) Scherer *et al.* demonstrated presumably reliable changes in 73 out of a total of 118 test measures.

The results of the original experiment appeared to support four generalized hypotheses which can be stated briefly as follows: (*a*) Immediately following lobotomy there is a general decline in mental efficiency reflected in lowered test scores at two weeks following lobotomy. Full recovery from the decline appears to take place by three months following lobotomy. At one year after the operation, a trend toward reduced efficiency is again noted. (*b*) Tests measuring strength of ego boundaries show a steady increase throughout the first postoperative year. (*c*) There appears to be greater sexual awareness and concern throughout all intervals during the first year following lobotomy. (*d*) Measures of rate of action and impulsivity suggest increased speed of reaction up to three months after lobotomy followed by a trend toward slower speed.

Certain methodological shortcomings in the original study, however, cast some doubts on these findings. First, since not all subjects could be tested on all measures at each occasion—either because of lack of cooperation or unavailability of the subjects due to illness, home visits, etc.—the sample utilized for the various comparisons of any single measure differed from one time interval to another. Second, it was found impossible to match experimental and control patients for preoperative performance on each test, although the groups were fairly well equated for such background data as age, education, length of hospitalization, level of cooperation, diagnostic classification, and certain behavioral characteristics (talkativeness, assaultiveness, hallucinatory experiences, etc.). Third, it has been demonstrated (8, 15) that if one gives a large battery of tests to two groups of people and then examines the data in order to find differences between the groups, a certain number of differences will be expected

---

[1] From the Veterans Administration Hospital, Northampton, Massachusetts. The authors wish to thank Drs. Robert L. Fortier and Frank Politzer, former postdoctoral trainees at this station, for assistance in testing patients; Dr. Robert W. Baker, Department of Psychology, Clark University, Lionel M. Ives, M.D., Chief, Professional Services, and Drs. Cesareo D. Peña, Arthur S. Tamkin, and Arnold Trehub, Staff Clinical Psychologists, for judging qualitative variables and critically reviewing this report; Messrs. Donald Isaac, Herbert Nickel, Joseph Mayer, and Alexander Stern, trainee Clinical Psychologists, for assistance in scoring quantitative measures and recording data.

Table 1

Age, Chronicity, and Diagnostic Groups for the Lobotomized and Control Patients in the First and the Second Experiment

Variable	Operated Patients		Control Patients	
	First	Second	First	Second
Number of Cases	22	15	22	10
Age, in years	33.9 (*SE* 1.3)	32.6 (*SE* 1.6)	30.4 (*SE* 1.0)	37.1 (*SE* 2.9)
Length of Hospitalization, in years	4.1 (*SE* 0.5)	6.2 (*SE* 1.7)	3.3 (*SE* 0.4)	7.9 (*SE* 2.4)
Diagnostic Group				
Catatonic	36.4%	53.3%	36.4%	50.0%
Paranoid	22.7%	13.3%	18.2%	20.0%
Hebephrenic	18.2%	20.0%	9.1%	20.0%
Mixed	22.7%	13.3%	36.4%	10.0%

to arise by the operation of chance factors alone. Thus, for example, of the 11 changes (each significant at the .10 level) in the direction of decreased mental efficiency two weeks following lobotomy, which Scherer reports in his Table 2 (12, p. 11), approximately four of them might be chance relationships.

For these reasons, the entire study was repeated on a different sample. The present study cannot be considered a replication of the original experiment (12) since important variations, to be discussed below, were made in the experimental procedure. However, it was thought that if the effects of lobotomy are fundamental and far-reaching enough to show up in spite of these variations, this information would certainly be useful.

Subjects

From a pool of 70 patients utilized in the Veterans Administration Central Office Research Project on Prefrontal Lobotomy (3), 43 subjects were sufficiently cooperative and willing to receive at least partial pretesting with the Northampton lobotomy battery. Of these, only 15 lobotomized and 10 control subjects were testable to any degree throughout the first year. These 25 individuals were selected as subjects for the study reported in this paper. The 15 lobotomized and 10 control patients were fairly well matched for age, diagnosis, chronicity, length of hospitalization, and a prognostic index based on past history and symptomatology present at the time of operation.[2] Table 1 gives the age, diagnostic classification, and chronicity for these patients as well as for those utilized in the first experiment.

An inspection of Table 1 indicates that although the first sample of operated patients was approximately the same age as the second sample, they had been hospitalized for a considerably shorter period of time. The control samples, however, differ with respect to age as well as hospitalization, the second sample being drawn from a much more chronic population. The difference was not unpredictable since most of the potential control patients had been operated on before the start of the present experiment. The differences in age and length of hospitalization undoubtedly influenced the results of psychological testing.

A second difference between the two experiments is that of the motivation of the patient-subjects. In the first study, 22 lobotomized and 22 control patients were given pretesting. Of these, 16 operated and 18 controls were willing and able to take the test battery one year later. In the second sample, however, much more attrition occurred. Of the 19 lobotomized and 24 control patients originally tested, only 15 operated and 10 controls took any one of the entire series of tests. It was the impression of the examiners, moreover, although no specific data exist to bear this out, that the original control sample appeared to be genuinely interested in taking the tests and that the second control sample was not.

The third difference between the samples used in the two studies is found in the type of operation undergone by the lobotomized patients. The study reported in the *Monograph* (12) was based on changes following an open Lyerly-Poppen approach while the present report dealt with changes following the McKenzie approach (nine patients) and the Scoville approach (six patients). The McKenzie approach substitutes a McKenzie leucotome for the spatula used in the Lyerly-Poppen approach; Scoville's bilateral cortical undercutting is a radically different operation.

---

[2] The prognostic index is derived from Schedules B-1 and B-2 of the Veterans Administration Central Office Research Project on Prefrontal Lobotomy (3).

Since inspection of the data showed no gross differences in the effects of the McKenzie and Scoville operations, the results were pooled for the present study. However, statements of nursing personnel suggest that the present sample of patients, regardless of the type of operation, showed much less immediate change in behavior in ward situations than did the patients in the first experiment. The discrepancy may be due, in part, to differences in the clinical status of the samples utilized in the two experiments, or, perhaps, to differences in the operative technique (5).

Procedure

All operated patients were individually tested prior to, and two weeks, three months, and one year following lobotomy. Control patients were individually tested at comparable intervals. This represents a further deviation from the procedure reported in the *Monograph* (12); in the earlier study, all patients were group-tested whenever possible. In the present study, however, since no more than two patients were operated on during any single week, group administration of tests was not possible. This probably influenced the results of testing, as has been suggested by Scherer and others (10, 13) in spite of an attempt on the part of the examiners to be as little involved in the situation as possible.

One hundred and six measures, derived from 27 psychological tests, were utilized in the lobotomy test battery (cf. 12, pp. 6–10, 17). Specific empirical predictions were derived from an analysis of the statistically significant changes in the first study as to the direction of the net change in the operated group for each of the test measures for each time interval under consideration: (*a*) pretest—two weeks posttest, (*b*) pretest—three months posttest, (*c*) pretest—one year posttest, (*d*) two weeks posttest—three months posttest, (*e*) two weeks posttest—one year posttest, and (*f*) three months posttest—one year posttest. Each prediction was individually tested by ascertaining the significance of the net change in means using formula 7.25 of Walker and Lev (14, p. 157). Since the earlier study suggested that the variance for the two samples could not be presumed to be equal, the degrees of freedom were estimated by formula 7.26 (14, p. 158). Dichotomized data were tested by using McNemar's net shift in proportion (6, pp. 81–82), the entries in the fourfold tables being the proportion of subjects possessing the attribute in question. One-tailed tests were used throughout, with the level of acceptance being set in advance at $\frac{1}{2}p = .05$.

Results

Two hundred and eleven specific predictions of net change for the operated group were empirically derived from the significant changes obtained in the original experiment (12).[3] Of these, only 15 were accepted at the .05 level of confidence, an outcome not significantly different from chance ($\chi^2 = 3.7027$, $df = 6$, $p > .50$). The outcome for each of the six time intervals under consideration was similar, as can be seen from Table 2. Those measures for which no prediction could be made, because they yielded no significant differences in the original experiment, also showed chance distributions of significant net changes. In view of the negative outcome, it was not felt worth while to examine in detail the status of the four generalized hypotheses of the *Monograph* (12). These hypotheses cannot be supported by the present study.

The conclusion reached from these results is that the two samples of operated patients utilized in this series of experiments were not affected in the same way. It seems that the hypothesized effects of lobotomy, as found in the first sample, are not so fundamental as to appear despite the changes in experimental procedure. The original conclusions derived from the *Monograph* (12) should be interpreted with care, therefore, since it is quite possible that the original conclusions are specific to the particular sample of patients under study and cannot be generalized to a universe of lobotomized patients who may have undergone a variety of operative procedures.

---

[3] Tables giving, for each measure, the number of cases, means, and the predicted and obtained net change in the operated sample for the periods (*a*) pretest—two weeks posttest, (*b*) pretest—three months posttest, (*c*) pretest—one year posttest, (*d*) two weeks posttest—three months posttest, (*e*) two weeks posttest—one year posttest, and (*f*) three months posttest—one year posttest, have been deposited with the American Documentation Institute. Order Document No. 4851 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $2.00 for microfilm or $3.75 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Table 2

Outcome of 211 Empirical Predictions as to the Effect of Prefrontal Lobotomy

Time Interval	Total Predictions	Accepted Predictions	Rejected Predictions	Chi Square ($df=1$)	p
Pre–Two Weeks					
Expected	39.0	1.95	37.05	0.5950	>.25
Obtained	39.0	3.00	36.00		
Pre–Three Months					
Expected	30.0	1.50	28.50	0.1753	>.50
Obtained	30.0	1.00	29.00		
Pre–One Year					
Expected	32.0	1.60	30.40	0.1052	>.50
Obtained	32.0	2.00	30.00		
Two Weeks–Three Months					
Expected	42.0	2.10	39.90	1.8094	>.10
Obtained	42.0	4.00	38.00		
Two Weeks–One Year					
Expected	33.0	1.65	31.35	0.0781	>.75
Obtained	33.0	2.00	31.00		
Three Months–One Year					
Expected	35.0	1.75	33.25	0.9397	>.25
Obtained	35.0	3.00	32.00		
Total					
Expected	211.0	10.55	200.45	3.7027 ($df=6$)	>.50
Obtained	211.0	15.00	196.00		

A study of changes in the present sample beyond the first year after lobotomy would be extremely valuable and was, indeed, planned by the present investigators. However, it is not being carried out since the current use of chemotherapeutic techniques with most of our operated and control patients might mask any differences which might arise from lobotomy as such.

Summary

A second study of the changes in psychological tests during the first year following prefrontal lobotomy has been carried out with a sample of 15 operated patients, all of whom were individually tested prior to, and two weeks, three months, and one year following prefrontal lobotomy. Ten equated control patients were individually tested at comparable intervals.

The present sample differed considerably from that used in the first study with respect to length of hospitalization, motivation, and type of operation. In addition, all patients were individually tested in the second experiment rather than being tested in a group.

The psychological test battery consisted of 27 tests from which 106 measures were derived. Two hundred and eleven specific predictions were derived from the results of an earlier study (12) as to the direction of net change in the operated group for each of six time intervals: (*a*) pretest—two weeks posttest, (*b*) pretest—three months posttest, (*c*) pretest—one year posttest, (*d*) two weeks posttest—three months posttest, (*e*) two weeks posttest—one year posttest, and (*f*) three months posttest—one year posttest. Each prediction was tested by ascertaining the statistical significance of the net change in means, or, in the case of dichotomized data, the net shift in proportion of the individuals possessing the attribute. One-tailed tests were used throughout the analysis with the level of acceptance being set in advance at the $\frac{1}{2}p = .05$ level.

The conclusions of the study were as follows: (*a*) only 15 out of 211 specific predictions were accepted ($\frac{1}{2}p = .05$), an outcome not significantly different from chance. (*b*) These results suggest that the hypothesized effects of lobotomy, as found in the first experiment, are not so far-reaching as to appear despite the variations in experimental procedure. (*c*) While a further study of changes beyond the first postoperative year might have been extremely valuable, it is not being carried out, since it would yield relatively little information as to the effect of lobotomy in the present sample because of the increasingly routine use of chemotherapeutic procedures with both lobotomized and nonlobotomized patients.

Received September 30, 1955.

References

1. Freeman, W., & Watts, J. W. Psychosurgery: an evaluation of 200 cases over seven years. *J. ment. Sci.*, 1944, **90**, 532–537.
2. Greenblatt, M., Arnot, R., & Solomon, H. C. (Eds.) *Studies in lobotomy*. New York: Grune & Stratton, 1950.
3. Jenkins, R. L., & Holsopple, J. Q. Criteria and experimental design for evaluating results of lobotomy. In *Psychiatric treatment* (*Proc. Ass. Res. nerv. ment. Dis.*, 1951, **31**) Baltimore: Williams & Wilkins, 1953. Pp. 319–327.
4. Kalinowsky, L. B., & Hoch, P. H. *Shock treatments, psychosurgery, and other somatic treatments in psychiatry*. (2nd Ed.) New York: Grune & Stratton, 1952.
5. King, H. E. Intellectual function. In F. Mettler (Ed.), *Selective partial ablation of the frontal cortex*. New York: Hoeber, 1949. Pp. 178–207.
6. McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.
7. Mettler, F. (Ed.) (Columbia Graystone Associates) *Selective partial ablation of the frontal cortex*. New York: Hoeber, 1949.
8. National Bureau of Standards, Applied Mathematics Series 6. *Tables of the binomial probability distribution*. Washington, D. C.: U. S. Government Printing Office, 1949.
9. Reitman, F. Observation on personality changes after leukotomy. *J. nerv. ment. Dis.*, 1947, **105**, 583–589.
10. Scherer, I. W. The psychological scores of mental patients in an individual and group testing situation. *J. clin. Psychol.*, 1949, **5**, 405–408.
11. Scherer, I. W., Winne, J. F., & Baker, R. W. Psychological changes over a three-year period following bilateral prefrontal lobotomy. *J. consult. Psychol.*, 1955, **19**, 291–298.
12. Scherer, I. W., Winne, J. F., Clancy, D. D., & Baker, R. W. Psychological changes during the first year following prefrontal lobotomy. *Psychol. Monogr.*, 1953, **67**, No. 7 (Whole No. 357).
13. Scherer, I. W., Winne, J. F., Page, H. A., & Lipton, H. An analysis of patient-examiner interaction with the Szondi pictures. *J. proj. Tech.*, 1952, **16**, 225–237.
14. Walker, Helen, & Lev, J. *Statistical inference*. New York: Holt, 1953.
15. Wilkinson, B. A statistical consideration in psychological research. *Psychol. Bull.*, 1951, **48**, 156–158.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

A Note on Research Methodology Used in Testing for Examiner Influence in Clinical Test Performance

Leon H. Levy
Indiana University

This note presents a brief methodological critique of research concerned with the problem of how much of the variance in clinical test performance may be accounted for by the personality characteristics of the examiner (E).

The extent of actual research concerned directly with this problem is quite limited and with the exception of Lord's study (2), each of these makes use of a *simple-randomized* design (1). This design is characterized by the random assignment of subjects (Ss) all drawn at random from the same parent population to various independently administered treatments (in this instance, Es). This is usually accomplished in one of two ways: (*a*) Ss are randomly assigned to each of several Es, (*b*) a random sample of cases already tested by each of several Es is pulled from case files. In this latter case it is assumed that there was no bias operating in the initial assignment of cases to each of the Es. In order to test the hypothesis of E influence on test performance, differences in mean scores between groups tested by various Es are then tested for significance. Any significant differences found are taken as evidence of E influence on test performance.

The limitations of the simple-randomized design may be summarized briefly as follows: Whatever control it accomplishes over sampling error is entirely contingent upon the randomization procedure used in initial case selection and assignment. Where case file data are used it is doubtful whether adequate randomization ever obtains; where Ss are assigned at random the possibility of intersubject differences still exists since randomization at best reduces the likelihood of bias but does not completely control it. A further limitation of this design is its failure to take interaction effects into account. Since it would seem highly unlikely that these do not exist in this area, data thus far accumulated cannot be adequately interpreted until they are known.

If case file data are to be used and one is willing to forego investigation and control of interaction effects, a *paired-replicates* design or *treatments-by-levels* design may be used. This would obviate to some extent the necessity of assuming initial random case assignment. The important requirement here would be that Ss in each of the E groups be matched on psychologically meaningful variables.

Where the experimenter has complete control over S selection and assignment, in addition to the paired-replicates design he might also use a *random replications* design (1). This design further reduces the likelihood of sampling error and permits some control over interaction effects through randomization.

Finally, although several other experimental designs could be used more legitimately than the simple-randomized design, it would seem that because of the likelihood and importance of interaction effects and in terms of the amount of information yielded, any design other than some type of factorial design must be considered profligate of time, effort, and data.

Received September 21, 1955.

References

1. Lindquist, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.
2. Lord, Edith. Experimentally induced variations in Rorschach performance. *Psychol. Monogr.,* 1950, **64**, No. 10 (Whole No. 316).

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

The Examiner as an Inhibiting Factor in Clinical Testing

Lewis Bernstein[1]

Veterans Administration Hospital, Denver, Colorado

This study attempted a simultaneous investigation of two factors in test administration, and their interaction. The factors studied were oral vs. written administration of the TAT, and the effects of the presence or absence of an examiner under each of these methods of administration.

In the two studies which have previously investigated oral vs. written administration of the TAT (3, 8), the oral condition was administered individually, and the written condition was administered in group form. For example, Eron and Ritter (3) administered the written condition to groups of six, with two subjects sharing one set of TAT cards. In addition to the unknown effects of the group situation, we have the further contaminating factor of the interaction of the two subjects using the same set of cards. Furthermore, the individual oral administration was according to standard directions, with no time limit, while the written group administration imposed a time limit of five minutes per card. It must be pointed out, however, that the conditions of Eron and Ritter's study were necessary since they were interested in determining whether the group method could economically be used in gathering normative data. However, the two forms of administration are not strictly comparable for determining the relative usefulness of written vs. oral stories in a clinical situation. It would appear, then, that there remains a need for comparing oral vs. written TAT stories, both administered individually.

As to the need for investigating the effect of the presence or absence of the examiner on test protocols, Sells (7) has reported that Rorschach responses appeared quite uninhibited in the group test (with relative absence of the examiner), but controlled in the individual administration. He goes on to state that ". . . one important hypothesis requiring careful study is that the presence of an examiner in the clinical test may inhibit the reporting of highly emotional content" (7, p. 25).

In this study, then, by using a 2 × 2 factorial design (1), we were able to investigate these two factors, as well as their interaction, which potentially could be of even more significance than the individual factors themselves.

Method

The subjects were 67 female college students—the entire sophomore class of the University of Colorado School of Nursing. The subjects were randomized into four groups: (*a*) Oral, examiner absent; (*b*) Oral, examiner present; (*c*) Written, examiner absent; (*d*) Written, examiner present. There were two examiners—one was the subjects' psychology instructor, the other completely unknown to any of the subjects prior to the experiment.[2]

Subjects in all conditions were shown into a private office and were told they would find directions for what they were to do and the necessary materials on the desk. On each desk was a set of TAT cards turned face down with card one on top. The cards were the third revision, and included all twenty pictures recommended by Murray (6) for use with adult females. The following directions for all subjects were typewritten on individual cards:

---

[1] The author is indebted to Dr. Richard H. Dana for his assistance in the collection of the data. Thanks are also due to the following persons who served as raters: Paul R. Binner, A. Frank Knotts, Mrs. Nancy M. Robinson, and Robert A. Spicer.

[2] The small number of examiners in this study does not fulfill the requirements of representative design, as suggested by Hammond (5), and the conclusions will therefore be limited to the two examiners in this study until the number of examiners has been increased.

> This is a task which you should find interesting. On the table before you is a set of pictures which you are to look at one at a time, and make up as dramatic a story as you can for each one. Tell what has led up to the event shown in the picture, describe what is happening at the moment, and then give the outcome. Since this is a test of how good your imagination can be, try to tell as much as possible about what the characters are feeling and thinking.
>
> When you get to Card No. 16, you will find it is a blank card. Imagine a picture for this card, and then go ahead and give a story about this imagined picture.
>
> Be sure to number your stories to correspond to the numbers on the cards. (Be sure to dictate the number of the card before dictating your story.)
>
> No further instructions will be given.

All subjects who were to be assigned to the oral conditions were given group instruction on the use of the dictaphone prior to being shown into the office in which they were to work. Each desk was equipped with either a pad of paper and pencils or a dictating machine, depending on the experimental condition. The only difference between the examiner-present and examiner-absent conditions was that in the examiner-present situation, one of the examiners sat across the desk from the subject throughout the session.

Each story was coded and typed on a duplicating stencil so that comparable sets of stories could be presented to the raters. Stories were rated on Eron, Terry, and Callahan's scale for emotional tone (4); on Eron's scale for outcomes (2); and on Terry's scale for level of response (9). Only one variable was rated at a time and all the stories for a given picture were rated together. All stories were independently rated by two judges. Pearson r reliability coefficients were .88 for emotional tone, .82 for outcomes, and .83 for level of response.

Results

Table 1 shows the differences in the means for the three scales under each of the four conditions of administration. For each scale, the examiner-absent condition yielded a significantly higher mean than the examiner-present condition, whether the stories were

Table 1

Differences in Means for All Scales

Variable	Oral, Examiner Absent ($N=17$)	Oral, Examiner Present ($N=16$)	Written, Examiner Absent ($N=18$)	Written, Examiner Present ($N=16$)	Mean Difference	t
Emotional Tone	−17.06	−11.13			5.93	3.01**
	−17.06		−17.33		.27	.15
		−11.13		−10.25	.88	.49
			−17.33	−10.25	7.08	4.19***
Outcomes	−8.71	+3.69			12.40	3.82***
	−8.71		−6.61		2.10	.67
		+3.69		+3.69	.00	.00
			−6.61	+3.69	10.30	4.68***
Level of Response	68.47	58.50			9.97	2.98**
	68.47		63.61		4.86	1.83
		58.50		60.00	1.50	.50
			63.61	60.00	3.61	1.60

Note.—For each of the examiner-present conditions the scores for the two examiners were combined since there were no significant differences on any scale between the two examiners.

** Significant at the .01 level of confidence.

*** Significant at the .001 level of confidence.

Table 2

Summary of Analysis of Variance

Variable	Source of Variation	*df*	Mean Square	*F*
Emotional Tone	Oral vs. Written	1	.55	.017
	Exam. abs. vs. Exam. pres.	1	708.99	21.193***
	Interaction	1	6.30	.188
	Within groups	63	33.45	
	Total	66		
Outcomes	Oral vs. Written	1	14.55	.228
	Exam. abs. vs. Exam. pres.	1	2140.59	33.541***
	Interaction	1	23.81	.373
	Within groups	63	63.82	
	Total	66		
Level of Response	Oral vs. Written	1	51.53	.761
	Exam. abs. vs. Exam. pres.	1	748.81	11.062**
	Interaction	1	177.67	2.625
	Within groups	63	67.69	
	Total	66		

Note.—In each case the hypothesis of homogeneity of variance is tenable.

** Significant at the .01 level of confidence.

*** Significant at the .001 level of confidence.

oral or written (with the exception of the written stories for level of response). There were no significant differences between oral and written stories for any of the scales, whether the examiner was present or absent.

Table 2 shows the results of the analysis of variance for each scale. In each case, the only significant *F* was for the examiner-absent vs. examiner-present condition. There was no significant interaction between the two testing factors studied.

Discussion

It has been demonstrated, within the limits of this experiment, that there is no appreciable difference between TAT stories given orally or written by the subjects in this experiment, as far as emotional tone, outcomes, and level of response are concerned. However, with the examiner absent the stories are sadder, have sadder outcomes, and show greater involvement on the part of the subject. These findings clearly support Sells' suggestion that the presence of an examiner may inhibit highly emotional content (7). It was noted by the examiners in the present study that subjects frequently looked up to see if the examiner was being attentive; and in the written condition, many subjects put their completed stories under the pad of paper, as if to assure that the examiner could not read them.

However, the mere physical absence of the examiner may not be the important variable. It seems more likely that what may be operative is the subject's expectancy for immediate evaluation. For example, Terry (8) reported a significantly higher level of response for individually administered TAT stories than for those given by written group-administration. Although the examiner was relatively absent from the group administration, she noted that "After completion of the first story all subjects were reminded of the directions if they had omitted any part of the story, such as the ending, or what had happened before to the characters" (8, p. 14). What may have happened here is that the subjects, after having had their first story evaluated by the examiner, expected that further evaluation might follow. On the other hand, this expectancy was apparently not created in the group administration of the Rorschach reported by Sells (7). In the present study, the examiner did not read or hear any of the stories given in the examiner-absent condition until the subject had left the examining room. In other words, it is tentatively proposed that the creation of an expectancy for immediate evaluation, even if further evaluation does not in fact occur, changes the testing-field situation for the subject from one of examiner-absent to examiner-present. It is planned to study this possibility in future experiments.

Summary

Sixty-seven female college sophomores were randomized into four groups for administration of the TAT under the following conditions: oral, examiner absent; oral, examiner present; written, examiner absent; and written, examiner present. The resulting 1,340 stories were rated for emotional tone, outcomes, and level of response. Analysis of

variance indicated that the only significant difference in the stories on each scale was a function of the presence or absence of the examiner, with the examiner-absent condition yielding higher means on each of the scales. There were no apparent differences between written and oral stories, nor was there any significant interaction between the two variables studied.

The data support the hypothesis that the presence of an examiner in a test situation acts as an inhibiting factor for strongly emotional material. It was suggested that the variable operative may not have been the mere physical presence or absence of the examiner, but rather an expectancy created in the subject for either immediate or more remote evaluation of his productions.

Received October 28, 1955.

References

1. Edwards, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950.
2. Eron, L. D. A normative study of the Thematic Apperception Test. *Psychol. Monogr.,* 1950, **64,** No. 9 (Whole No. 315).
3. Eron, L. D., & Ritter, Anne M. A comparison of two methods of administration of the Thematic Apperception Test. *J. consult. Psychol.,* 1951, **15,** 55–61.
4. Eron, L. D., Terry, Dorothy, & Callahan, R. The use of rating scales for emotional tone of TAT stories. *J. consult. Psychol.,* 1950, **14,** 473–478.
5. Hammond, K. R. Representative vs. systematic design in clinical psychology. *Psychol. Bull.,* 1954, **51,** 150–159.
6. Murray, H. A. *Manual for the Thematic Apperception Test.* Cambridge, Mass.: Harvard Univer. Press, 1943.
7. Sells, S. B. Problems of criteria and validity in diagnosis and therapy. *J. clin. Psychol.,* 1952, **8,** 23–28.
8. Terry, Dorothy. An analysis by the use of rating scales of Thematic Apperception Test protocols obtained under different conditions of test administration. Unpublished doctor's dissertation, Univer. of Wisconsin, 1950.
9. Terry, Dorothy. The use of a rating scale of level of response in TAT stories. *J. abnorm. soc. Psychol.,* 1952, **47,** 507–511.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

An Effect of the Situation on Psychological Testing

E. Jerry Phares
Kansas State College

and Julian B. Rotter
The Ohio State University

A number of recent studies have begun to emphasize the role of situational variables in psychological testing (1, 2, 3, 6). Such variables include the examiner, the test instructions and the set of the examinee, and the place of testing. Although these studies have effectively demonstrated the importance of such variables, they have not provided a basis for the regular or systematic prediction of the nature and amount of variation in performance due to such variables.

Social learning theory provides the hypothesis that one useful way to characterize situational variables and their effect on behavior is in terms of the expectancies they arouse for reinforcement. The expectancy that a particular reinforcement will follow from a particular behavior is in part a function of the specific situation and the expectancy that a particular reinforcement will be followed by later specific reinforcements is also a function of the situation. More detailed descriptions of these relationships have been made by Rotter (4, 5).

Simply stated, situations provide cues for expected reinforcements, both immediate and delayed. If behavior is directed toward maximum gratification (or minimum punishment) we should be able to predict that behaviors characteristically directed toward particular reinforcements (or away from in the case of negative reinforcements) would be most likely to occur in situations which we can identify on the basis of previous experience as most likely to provide the reinforcements in question.

Although this is a simple, common-sense proposition, it has not been put to work systematically in attempting to predict the effect of such variables as the personality or sex of the examiner or the time and place of testing for either clinical or experimental testing. Many failures to duplicate test results both in clinical work and in experimental investigations have been ascribed to unreliability in the subjects or test instruments rather than to differences in the testing situation.

The present study is an attempt to test the hypothesis that the preference values of a list of reinforcements will differ predictively depending upon the situation in which they are obtained when random groups of subjects drawn from the same population are used.

More specifically, it was hypothesized that differences greater than chance would be found in the rankings of three kinds of rewards—judged to be academic achievement rewards, athletic achievement rewards, and manual skill rewards—when obtained in an English class, a gymnasium, and a woodworking glass by the respective instructors in those classes. It was hypothesized that academic achievement reinforcements would be rated by subjects to have higher value in the English class, athletic achievement in the gymnasium, and manual and mechanical skill achievement rewards in the woodworking class. This study was also undertaken to explore one methodology for determining the significance of the psychological situation.

Procedure

It was important that the items provided for ranking should all be of approximately equal value; this value being moderate. Reinforcements of moderate value would be free to move either up or down in value as the situation changed. Three lists of reinforcements (each containing 10 Athletic, 10 Academic, and 10 Manual reinforcements in random order) were constructed. Each reinforcement had previously been unanimously categorized as exclusively either Manual (*M*), Academic (*Ac*), or Athletic (*At*) by five judges. Three groups of male seventh and eighth graders (comprising a total of 99 subjects), were asked to categorize each item under one of five headings: (*1*) Like hardly at all; (*2*) Like some; (*3*) Like pretty well; (*4*) Like a lot; and (*5*) Like very much. The subjects were thereby to assign a value to each reinforcement. Since the expectancy of the occurrence of a reinforcement is a determiner of the value of that reinforcement, this expectancy was held constant for each reinforcement by this statement prefacing each list: "Imagine that you could do, become, or have all the things in the sentences below. . . ."

The data from the administration of these lists were analyzed as follows: for each item, the number of subjects placing it under each heading was tabulated; each heading was assigned a numerical value of from 1 to 5—thus, "Like hardly at all" received a value of *1* and "Like very much" received a value of *5*; a value for each item was obtained by multiplying the number of cases under each heading by the value of that heading; then, for each item, the five values were summed and divided by the number of subjects who dealt with that item.

From these data it was possible to get three matched groups of *At*, *Ac*, and *M* reinforcement items, each group having six items and the following mean values: Athletic 3.06; Academic 3.04; Manual 3.05. These figures satisfy the criterion of moderate value.

These 18 items were then randomly arranged into the list to be used in the final research, and contained such representative items as, "Be praised by the teacher for writing a good book report," "Have your friends admire you for getting good grades," "Win a wrestling match with a friend," "Know more about baseball than the other guys," "Get a tool set for Christmas," "Win a prize for building the best magazine stand in school."

Subjects for the final phase of the experiment were 78 eighth-grade and 72 seventh-grade boys, all from the same school. Each of the three situations, English classes (Academic), gym classes (Athletic), and woodworking classes (Manual), were required classes for all male students of the school. Fifty different subjects ranked the list of reinforcement items in each situation. In all three situations, the ranking procedure was not initiated until the classes had been underway for at least ten minutes, to help ensure the representativeness of the situations. In the English classes, the ranking was done with the subjects seated at their desks; in the gym classes the subjects were clad in gym clothes and seated on bleachers; in the woodworking classes the subjects did the ranking at their benches. In every case, the particular class instructor administered the list. Since all were required courses it was felt that the samples of subjects in each situation were large enough to ensure random selection of seventh and eighth graders for this school population.

Results

The mean ranks for the three groups of reinforcement in each situation are shown in Table 1. The lower the score the higher the preference.

Reliability coefficients for the list of reinforcements were computed by the odd-even method and corrected by the Brown-Spearman formula. They were .79 for Athletic reinforcements in the Athletic situation; .91 for Academic reinforcements in the Academic situation; .83 for Manual reinforcements in the Manual situation.

Table 1

Comparison of Mean Ranks of Reinforcement Items in Three Situations

Items	Situation			Mean of 3 situations
	Academic	Athletic	Manual	
Academic	9.54	10.21	9.26	9.67
Athletic	9.83	8.40	9.90	9.38
Manual	9.12	9.82	9.28	9.41

Table 2

Significance of Differences of Means of Pairs of Reinforcements for Two Situations

Situation and reinforcements	Difference	p^*
Academic vs. Athletic	2.10	$<.001$
Athletic vs. Manual	2.04	$<.001$
Manual vs. Academic	.44	$>.10$

* p values based on a one-tailed test of significance, since direction as well as change was predicted.

Since the differences between mean rankings were slight, difference scores were used to test the hypothesis by calculating the significance of (*a*) the difference between mean ranks of the Academic and Athletic reinforcements in the Academic situation as compared with that difference in the Athletic situation; (*b*) the difference between mean ranks of the Manual and Athletic reinforcements in the Manual situation as compared with that difference in the Athletic situation; and (*c*) the difference between mean ranks of the Manual and Academic reinforcements in the Manual situation as compared with that difference in the Academic situation. The data are shown in Table 2.

It will be seen that significant results were found in the case of (*a*), differences between the Athletic and Academic reinforcements in these two situations and (*b*), differences between the Athletic and Manual reinforcements in these two situations. No significant differences, however, were obtained between the rankings of (*c*), Academic and Manual reinforcements in these two situations.

This latter failure to demonstrate differences may be partially explained, it is felt, on the basis of the greater similarity of these two situations (*M* and *Ac*) as compared with the Athletic situation, which contained rather distinctive cues. An additional explanation after a *post hoc* analysis of the Manual items, such as, "Have your dad let you tinker around with his car," and "Work on your bicycle after school," showed that several of them had little in common with what went on in the woodworking classes and did not as it was presumed, involve similar skills to those rewarded in a woodworking class.

Discussion

The results show that two of the three predicted differences occurred to a significant degree. Although the specific test responses studied may not be of great significance in themselves, the study suggests that such relatively unimportant variables as the physical setting of the testing and the subjects' presumed characterization of the examiner (English teacher, gym teacher, etc.) may generally have demonstrable effects on test results. It seems likely that similar effects may occur if the subject is a hospital patient and sees the examiner as some one in authority or not, or if in a typical experiment the subjects presume that the experimenter is an instructor or a graduate student, or if they anticipate that the examiner is one who can provide them with help, or appears to appreciate social skills more than intellectual ones, etc. In other words, the present study lends some support to the hypothesis that in both clinical and experimental testing, the group or individual subject's behavior must be appraised in light of the reinforcements they expect to be more likely to occur in the particular situation provided by the physical setting and the examiner, as well as by the test instructions.

Received October 12, 1955.

References

1. Gibby, R. G., Miller, D. R., & Walker, E. L. The examiner's influence on the Rorschach protocol. *J. consult. Psychol.* 1953, **17**, 425–428.
2. Lord, Edith. Experimentally induced variations in Rorschach performance. *Psychol. Monogr.*, 1950, **64**, No. 10 (Whole No. 316).
3. Luchins, A. S. Situational and attitudinal influences on Rorschach responses. *Amer. J. Psychiat.*, 1947, **103**, 780–784.
4. Rotter, J. B. *Social learning and clinical psychology*. New York: Prentice-Hall, 1954.
5. Rotter, J. B. The role of the psychological situation in determining the direction of human behavior. In M. R. Jones (Ed.), *Nebraska symposium on motivation, 1955*. Lincoln: Univer. of Nebraska Press, 1955. Pp. 245–269.
6. Weisskopf, Edith A., & Dieppa, J. J. Experimentally induced faking of TAT responses. *J. consult. Psychol.*, 1951, **15**, 469–474.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

The Relationship of the Progressive Matrices (1938) and the Columbia Mental Maturity Scale to the WISC

Ernest S. Barratt
University of Delaware

Since its clinical inception in 1949, the Wechsler Intelligence Scale for Children (WISC) (5) has become one of the leading tests for measuring the general intellectual level of children. Two other scales, the Progressive Matrices (PM) (4) and the Columbia Mental Maturity Scale (CMMS) (1), are becoming increasingly popular measuring instruments of the general reasoning ability of children.

The CMMS was designed to provide "an estimate of intellectual ability in the age range 3 to 12 years" (1, p. 1). The CMMS is especially useful for testing children with speech difficulties and other handicaps.

The PM items appear to measure factorially complex spatial ability and reasoning by analogy. Although the author of the PM does not describe it as a test of *g* (after Spearman), the manual reports that it had a *g* saturation in one study of .82 (4, p. 2). In a recent normative study of the PM (1947 revision), the authors concluded that the PM should be considered "as a test of fairly complex intellectual reasoning processes" (2, p. 142) rather than a test of nonverbal reasoning ability.

Since the entire WISC cannot always be administered because of time limitations or physical handicaps of the subject, it would be helpful to know to what extent the PM or CMMS could be substituted for it. The purpose of the present study is to note the relationship of the WISC to the CMMS and PM.

Procedure

Seventy children, the entire fourth-grade enrollment in the Medill Elementary School, Newark, Delaware, were given the WISC, PM (1938), and CMMS as individually administered tests. The directions in the respective manuals for administering the tests were followed, with one exception, i.e., in administering the PM, the test was stopped when a subject missed six consecutive items. The age range of the children was from 9.2 years to 10.1 years. The children came from both rural and urban homes. There were 26 boys and 34 girls in the study. The 1938 edition of the PM was used since it has been the author's experience that in testing children of 8 years and older, sufficient variance is obtained with the 1938 edition to obviate substituting the 1947 revision of the PM.

Results and Discussion

Table 1 contains the means and standard deviations (*SD*) for the various tests used in this study. The WISC means and *SD*s are in

Table 1

Means and Standard Deviations of the WISC, PM, and CMMS

Test	Mean	*SD*
WISC*		
Total	108.92	18.02
Verbal	55.14	11.11
Performance	53.78	8.68
Information	11.04	2.98
Comprehension	10.30	3.68
Arithmetic	12.30	2.66
Similarities	11.59	3.13
Vocabulary	10.81	3.78
Digit Span	10.17	2.36
Picture Completion	10.63	3.31
Picture Arrangement	11.04	3.14
Block Design	10.82	3.44
Object Assembly	10.94	2.46
Coding	10.69	2.68
Mazes	10.24	2.60
PM	24.11	7.33
CMMS	78.37	8.18

* All WISC scores are Scaled Score Units (5).

scaled score units to facilitate comparing the present sample with the original standardization sample (mean = 10, *SD* = 3 for WISC standardization subtests). Although the current WISC subtest means are slightly above average compared to the original WISC standardization sample, none of the deviations are significantly different from 10. The *t* ratio for the largest deviation from 10 (arithmetic, 12.30) is approximately 1.1. The WISC subtest *SD*s range from 2.36 (digit span) to 3.68 (comprehension), with the average of the *SD*s being 3.02. The subjects did slightly better on the verbal scales than on the performance scales but were less variable on the performance scales.

The PM mean of 24.11 corresponds to a chronological age (CA) of 9.5 years in the standardization sample (4, p. 12). The PM range in the current study is 3 to 43; the *SD* is 7.33. Using the 1947 Colored PM, Green and Ewert (2, p. 140) obtained a range of 9 to 35 and Martin and Wiechers (3, p. 144) obtained a range of 8 to 35 with comparable age groups. It is this author's opinion that the 1938 PM provides a better differentiation of ability at the extremes than does the 1947 Revision of the PM with children above a CA of 8 years.

The CMMS mean in this study corresponds to a mental age of 12 years, and an IQ of 144 in the standardization sample (1, p. 8). Although the authors of the CMMS indicate that the IQs of the standardization group on the CMMS correspond to their Stanford-Binet IQs (1, p. 11), the CMMS has consistently yielded higher IQs than either the Binet or WISC in the University of Delaware Laboratory Clinic.

Table 2 contains the uncorrected correlations of the PM and CMMS with the WISC. The PM was significantly related to the WISC total, verbal, and performance scores and to all of the WISC subtests except Comprehension. Considering the rank order of the PM-WISC subtest *r*s, the PM correlates highest with those tests involving spatial reasoning (block designs), verbal reasoning of a more or less abstract nature (similarities), and acquired knowledge (information and vocabulary). The PM has the lowest correlations with tests that involve reasoning in a cultural context (comprehension, picture arrangement) and tests involving perceptual speed and psychomotor tasks (coding, mazes).

Table 2

The Relationship of the Wechsler Intelligence Scale for Children to the PM and CMMS

WISC	CMMS	PM
Total	.606*	.754*
Verbal	.559*	.692*
Performance	.478*	.699*
Information	.129	.585*
Comprehension	.330*	.075
Arithmetic	.680*	.540*
Similarities	.425*	.589*
Vocabulary	.452*	.561*
Digit Span	.272**	.419*
Picture Completion	.377*	.415*
Picture Arrangement	.484*	.300*
Block Design	.473*	.601*
Object Assembly	.285**	.388*
Coding	.442*	.332*
Mazes	.642*	.323*
CMMS		.576*

* 1% level of confidence.
** 5% level of confidence.

The CMMS was also significantly related to the WISC total, verbal, and performance scores and to all of the WISC subtests except information. Ranking the CMMS-WISC subtest *r*s, it is difficult to interpret some of the *r*s. The CMMS had a higher *r* with arithmetic than any other subtest; since numerical concepts (counting, relative length, relative size, etc.) are involved in many of the CMMS items, this is not surprising. It is difficult, however, to account for the high *r* between the CMMS and mazes. The author could think of no logical connection between the problem-solving behavior required for solving the CMMS and maze items that could not also apply to some of the other, lower subtest CMMS *r*s.

The PM correlated higher than the CMMS with the WISC total, performance, and verbal scores. About 57 per cent of the WISC total variance was in common with the PM, while only 36 per cent of the WISC total variance was in common with the CMMS. At the age level of the subjects in the current study, the PM appears to be a more appropriate substitution than the CMMS for the WISC total score.

Summary and Conclusions

The Columbia Mental Maturity Scale (CMMS) and the Progressive Matrices (1938) (PM) were related to the WISC. Seventy fourth-grade children were tested. Both the PM and the CMMS were significantly related to the WISC total, verbal, and performance scores. The PM had more variance in common with the WISC total score than the CMMS. The relationship of the PM and CMMS to the WISC subtests was also discussed.

Received October 11, 1955.

References

1. Burgemeister, Bessie B., Blum, Lucille H., & Lorge, I. *Manual for Columbia Mental Maturity Scale*. Yonkers, N. Y.: World Book Co., 1954.
2. Green, M. W., & Ewert, Josephine C. Normative data on Progressive Matrices. *J. consult. Psychol.*, 1955, **19**, 139–142.
3. Martin, A. W., & Wiechers, J. E. Raven's Colored Progressive Matrices and the Wechsler Intelligence Scale for Children, *J. consult. Psychol.*, 1954, **18**, 143–144.
4. Raven, J. C. *Guide to using Progressive Matrices.* London: H. K. Lewis, 1938.
5. Wechsler, D. *Wechsler Intelligence Scale for Children*. New York: Psychological Corp., 1949.

Journal of Consulting Psychology
Vol. 20. No. 4, 1956

Pre-Illness Intelligence of Mental Hospital Patients[1,2]

Charles F. Mason
Purdue University[3]

In an undergraduate course in which the writer was enrolled the question was once raised as to whether there is any relationship between intelligence and psychological adjustment. The professor replied that no one could actually answer the question because relevant studies had always had to measure intelligence of maladjusted persons after they had become ill enough to appear in a hospital or clinic. It was speculated at the time that the question might ultimately be answered as an outgrowth of some plan of mass mental measurement such as the testing of all draftees during World War II. A few years later, when the writer was a VA trainee in clinical psychology, this point was recalled and it was realized that he was in a favorable position to obtain pre-illness (or at any rate pre-acute-illness) intelligence scores of a large number of psychiatric patients.

Since psychiatric patients can be considered as persons who have failed to make adequate psychological adjustments, by examining their Army General Classification Test scores (the AGCT is described in references 1, 8, and 9) we would be able to make inferences about relationship between intelligence and psychological adjustment in the same manner that we might study the relationship between intelligence and scholarship by examining the IQ scores of students who made *F*'s in courses. It would not be the whole story, but would be a beginning. Further, information about how persons who are now psychiatric patients performed on an intelligence test given before they became patients is important to the body of psychological knowledge. The performance of patients in different diagnostic groups would also be important to ascertain.

Investigation as to the practicality of the project disclosed that VA installations could obtain the AGCT scores of patients by sending a form to the Army. To fill out these forms for a large number of cases certainly posed a long and monotonous task, but it was decided to be worth the effort, and the study was undertaken. The project was carried out at the VA neuropsychiatric hospital at Marion, Indiana.

Planning the Study

In the literature pertinent to the question of relationship between intelligence and psychological adjustment, three general types of hypotheses were found to predominate. These were: (*a*) that there is a direct, positive relationship between intelligence and psychological adjustment; (*b*) that a particular intelligence level is optimal for adjustment in our society; (*c*) that patients in certain psychiatric diagnostic groups are brighter than those in others. These hypotheses led to the choice of chi square as the appropriate statistic for the present study.

If intelligence and adjustment are positively related, comparing AGCT scores of mental patients to those of controls by chi square should show that this maladjusted group had scored lower than the normals. In addition, inspection of the chi-square table would reveal any range of intelligence with low representation among patients, in accordance with the hypothesis of an optimal range

---

[1] This study is based upon a dissertation submitted by the writer in partial fulfillment of the requirements for the Ph.D. degree in psychology. The writer is indebted to Professor John M. Hadley for his critical and helpful service.

[2] From the Veterans Administration Hospital, Marion, Indiana.

[3] Now at the Veterans Administration Hospital, Long Beach, California.

of intelligence for adjustment. Each diagnostic group could be compared separately to controls by chi square to determine whether the group had scored significantly low or high on the AGCT.

The literature provided further guidelines in the plan of our study. That careful attention to design was necessary was shown by Lorge's review of studies of intelligence and adjustment, published in 1940 (6). He found a range of correlation coefficients from minus .49 to plus .70, with a median of .04. Of these findings he says that the range is so "extraordinary that anybody can make any statement." Such diverse findings must come from differences in experimental design. Spragg (7) and Williams (11) have discussed typical defects of design which apply to many relevant studies. Small sample and inadequate controls are frequent limitations, and fortunately bountiful and adequate data were available for the present project.

Collecting the Data

Nearly ideal control data were available in Davenport's distribution of AGCT scores of 290,163 Fifth Service Command Army inductees, so geographical area could be held constant. However, one difficulty was encountered. Davenport's distribution is broken down into only five class intervals. To make our test as sensitive as possible to the occurrence of an "optimal range," a breakdown into five-point class intervals was desired.[4] To accomplish this, Davenport's percentage values were further broken down according to values from a distribution of AGCT scores for the total Army.[5] Careful consideration was given to this procedure, which robs our design of elegance, but it was not judged to enter any practical bias.

The experimental sample consisted of 510 AGCT scores of World War II, male, Army veterans, who were or had been hospitalized at VA, Marion, and diagnosed by a staff of two or more psychiatrists as suffering from functional psychiatric disorders. The sample was obtained by drawing randomly from the hospital's register and rejecting cases which did not meet the criteria.

In discussion of the project with colleagues the writer found a very pertinent question frequently raised. Suppose some of the patients' AGCT scores were lowered by *acute* illness at the time of taking the test? A substudy designed to check this hypothesis was made. By analysis of variance it tested the assumption that subjects who had been hospitalized for schizophrenia shortly after taking the test would tend to have scored lower than schizophrenics hospitalized after a longer intervening period of duty. This investigation yielded no support for the hypothesis, so there was no further screening of the data.

Results

Figure 1 shows graphically the distributions for the experimental sample of AGCT scores of 510 patients and the control distribution based on scores of 290,163 Fifth Service Command inductees. Table 1 presents the results of comparing these two distributions by chi square as well as comparisons to the control data of distributions of AGCT scores for various diagnostic subgroups.

The patients were not found to have scored lower on the AGCT to an extent that is statistically significant. Schizophrenics, as a total group, scored significantly low. However, this result is misleading, since study of schizophrenic types reveals that the paranoid and catatonic types did not score low. In view of the often encountered statement that it requires high intelligence to be paranoid it might be well to underline the fact that the paranoid schizophrenic group showed no tendency whatever toward higher scores than the control group. The manic-depressive group, though including only 25 cases, showed a statistically significant high scoring trend.

Discussion

1. Two conclusions may be drawn from the comparison of the total sample of patients'

---

[4] Larger class intervals were used in the studies of diagnostic groups. These distributions were "telescoped" in order to maintain a minimum theoretical frequency of ten.

[5] U. S. Army, The Adjutant General's Office, Personnel Research Section. *Estimated percent distribution of AGCT scores* (Based on 9,757,583 AGCT scores by AGCT grade). A table prepared in July, 1947, and obtained from the Adjutant General's Office.

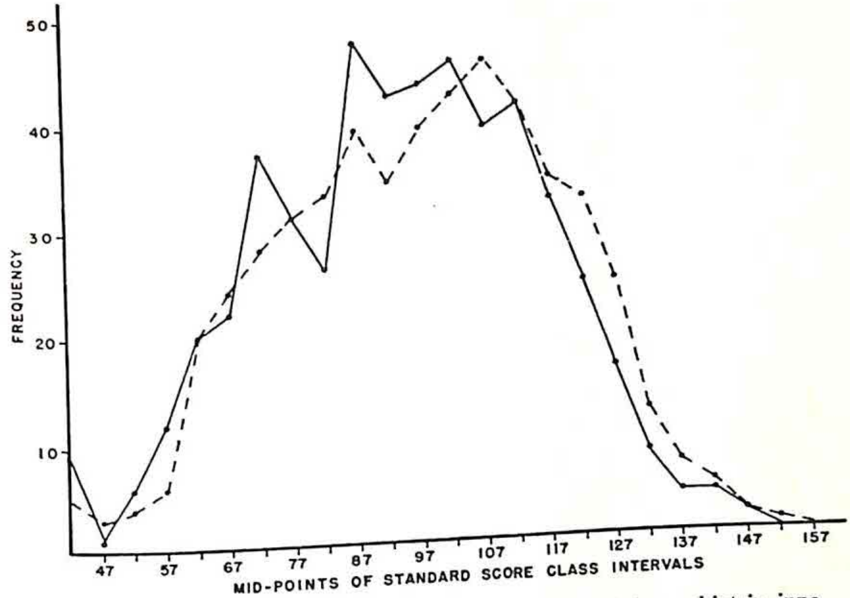


Fig. 1. Distribution of AGCT standard scores for 510 psychiatric inpatients (solid) and control distribution based on data for 290,163 Fifth Service Command inductees (broken).

scores to the control data: First, no significant relationship between intelligence and adjustment is shown. One might remark that we have shown 86 chances in 100 that "mentally ill" persons tend to have scored lower than the general population. However, the facts that this sample represents an extreme of maladjustment and that most of its eccentricity is due to certain schizophrenic types suggest to the writer that, if a single continuum, "maladjustment," were extended into the total U. S. population, any relationship to be found between this and intelligence would be negligible. Second, it is concluded that no particular level of intelligence is revealed as optimal for psychological adjustment (no restricted range along the intelligence continuum showed a low incidence in this sample of scores of maladjusted persons). It should be stated that this latter point has no bearing on the hypothesis of Hollingworth (4). Her hypothesis involves intelligence levels above IQ 130, which this study could not adequately sample. Thus our findings relative

Table 1

Chi-square Results: Comparisons to Control Data of AGCT Data for All Patients and for Diagnostic Subgroups

Group	*N*	*df*	Chi square	*p*	Trend
All patients	510	16	22.32	.14	
All schizophrenics	368	16	32.60	.02	Low scores
Paranoid	66	3	3.10	.39	
Catatonic	126	7	10.61	.19	
Hebephrenic	122	7	17.55	.02	Low scores
Simple	33	2	12.46	.002	Low scores
Other	21	1	4.77	.05	Low scores
Nonschizophrenics	188	12	7.78	.79	
Alcoholism	29	1	0.61	.90	
Neurosis	79	5	1.93	.20	
Manic-depressive	25	1	17.65	.001	High scores
Character disorder	55	3	4.21	.25	

to the first two hypotheses taken from the literature are negative.

With respect to the third hypothesis our findings are positive. From the investigation of intelligence of patients by diagnostic groups is drawn the conclusion that intelligence can be significantly related to diagnosis. Manic-depressive psychosis is associated with high AGCT score. Schizophrenia, other than paranoid and catatonic types, is associated with low AGCT score. This is in accord with the findings of the sociologic studies of Duval (2), Faris and Dunham (3), and Weiss (10), who found schizophrenia to have a higher incidence in lower socioeconomic groups, manic-depressive psychosis in upper socioeconomic groups.

Theoretical speculations to explain the findings for different diagnostic groups are interesting to make. However, this study was undertaken as an empirical rather than a theoretical one and provides little or nothing for evaluation of hypotheses the speculator might formulate. The writer's favored hypothesis about the low scoring tendency of some schizophrenics is that it is associated with personality disorganization of long standing, existing prior to acute illness and tending to have rendered the persons "functionally unintelligent." One might hypothesize a sharp cleavage between the scores of schizophrenics divided into reactive vs. process groups as discussed by Kantor, Wallner, and Winder (5). The trend of paranoid and catatonic schizophrenics scoring normally with all other schizophrenic groups scoring lower might become more pronounced through a resorting of atypical cases were the dichotomy of these authors substituted for the Kraepelinian typology. No speculations about the findings for manic-depressives are volunteered.

Summary

This study involved drawing a large sample of male, World War II, Army veterans from the files of the Veterans Administration neuropsychiatric hospital at Marion, Indiana. Army General Classification Test scores were obtained from the Army for 510 cases, this test having been taken prior to the onset of acute illness. The distributions of scores for this total group and for various diagnostic subgroups were compared to a control distribution based on AGCT data from 290,163 Fifth Service Command Army inductees. The findings were:

1. No definite relationship was found between intelligence, as measured by AGCT, and maladjustment as the latter was represented by the total sample.
2. Within the range of intelligence sampled by this study no particular level of intelligence appeared as optimal for psychological adjustment.
3. Relationships between intelligence, as measured by AGCT, and diagnostic classification were indicated: (*a*) Schizophrenics, other than paranoid and catatonic types, tended to have obtained low AGCT scores. (*b*) Manic-depressive psychotics tended to have obtained high AGCT scores.

Received September 12, 1955.

References

1. Davenport, R. K. Implications of military selection and classification in relation to universal military training. *J. Negro Educ.*, 1946, **15**, 585–594.
2. Duval, M. Psychoses in officers of World War II. *War Med.*, 1944, **5**, 1–5.
3. Faris, R. E., & Dunham, H. W. *Mental disorder in urban areas.* Chicago: Univer. of Chicago Press, 1939.
4. Hollingworth, Leta S. Intelligence as an element in personality. *Yearb. nat. Soc. Stud. Educ.*, 1940, **39** (I), 271–275.
5. Kantor, R. E., Wallner, J. M., & Winder, C. L. Process and reactive schizophrenia. *J. consult. Psychol.*, 1953, **17**, 157–162.
6. Lorge, I. Intelligence: its nature and nurture. *Yearb. nat. Soc. Stud. Educ.*, 1940, **39** (I), 275–277.
7. Spragg, S. D. Relations between intelligence and morbid addiction. *Yearb. nat. Soc. Stud. Educ.*, 1940, **39** (I), 285–290.
8. U. S. Army, The Adjutant General's Office, Personnel Research Section. The Army General Classification Test. *Psychol. Bull.*, 1945, **42**, 760–768.
9. U. S. Army, The Adjutant General's Office, Personnel Research Section. The Army General Classification Test, with special reference to the construction and standardization of Forms 1a and 1b. *J. educ. Psychol.*, 1947, **38**, 385–420.
10. Weiss, I. I. Psychoses in military prisoners. *J. clin. Psychopath.*, 1947, **8**, 801–815.
11. Williams, H. N. Intelligence and delinquency. *Yearb. nat. Soc. Stud. Educ.*, 1940, **39** (I), 291–294.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

The Revalidation of Diagnostic Tests for Alcohol Addiction[1]

Donal G. Murphy
Fordham University[2]

A review of reports extending through approximately 20 years, during which a large number of testing techniques have been applied in efforts to uncover the personality correlates of alcohol addiction, must lead to the conclusion that results have been most meager. Surveys, such as that of Sutherland, Schroeder, and Tordella (7) dealing with researches reported prior to 1950, have generally concluded that there is little objective evidence indicating that any particular personality type is predisposed to alcoholism. A search of the literature since 1950 has indicated no marked change in the status of this field of inquiry.

The Inductive Approach

Noting the discouraging results obtained in the application of standard testing methods, some investigators have approached the problem inductively; in such studies, the effort has been made to develop tests which would reliably identify alcohol addicts. The results of such tests, when properly validated, were expected to produce data the analysis of which might lead to a clearer picture of the psychic substrates underlying this addiction.

Among the efforts in this direction in recent years, the work of Manson (2, 4) in producing two diagnostic questionnaires yielded initial results of considerable interest. Manson's aim was to clarify by psychometric differentiation specific emotional, attitudinal, and behavioral patterns in alcohol addicts which differed significantly from such patterns among nonalcoholics. His Manson Evaluation (3) was an attempt to assess certain personality traits by questions relating to habitual feelings, reactions to persons and situations, goals and aspirations, physical states, and emotional trends. In the Alcadd Test (5), on the other hand, questions were openly directed toward elucidation of drinking habits and of affective responses to drinking by the subject and by others. Both tests were validated on several hundred alcoholic and nonalcoholic subjects of both sexes and the author claimed high validity, reliability, and predictive power for the two instruments. Barillas (1) confirmed Manson's results with the Alcadd Test only, administering this test to 200 male subjects including alcohol addicts, social drinkers and abstainers, and finding that it did reliably differentiate members of the three groups.

The purpose of the present research (6) was the revalidation of both tests, making use of female subjects who met carefully defined and applied criteria.

Method

While this revalidation might have been attempted using only "alcoholic" and "nonalcoholic" classifications of subjects, it was apparent that results of possibly greater significance might be obtained if specific groups lying along a postulated continuum running from no use of alcohol to plainly excessive use thereof could be tested. A differentiation within the alcoholic group between active addicts and members of Alcoholics Anonymous suggested itself, and it further appeared that there might be significant differences between women who customarily use alcoholic beverages and those who rarely or never do so. Four test categories were set up: active alcoholics, AA members, social drinkers, and abstainers. Since the Manson Evaluation was designed to reflect personality correlates underlying the use of alcohol, it was expected that the group mean for active alcoholics would be highest (most deviant), followed by those for AA members, social drinkers, and abstainers in that order. In the case of the Alcadd Test, it was decided that AA members would be asked to respond on the basis of habits antedating entrance into Alcoholics Anonymous, since it seemed of little value to obtain a drinking habits report from a group of professed abstainers. In this case, it was expected that means for both alcoholic groups would be high, with those for the other groups progressively smaller.

---

[1] Based upon a dissertation submitted in partial fulfillment of the requirements for the M.A. degree, Fordham University. The writer wishes to express his gratitude to William C. Bier, S.J., for his advice and helpful criticism.

[2] Now at Teachers College, Columbia University.

Criteria applied in the selection of test subjects were carefully worked out so as to differentiate distinct samples. In the case of the active alcoholics, the basic requirement was that the subject be under treatment in a hospital, sanitarium, or clinic specifically for alcoholism. Added conditions were that the subject be diagnosed psychiatrically as nonpsychotic and that last previous consumption of alcohol had occurred 5 to 14 days prior to testing. The time period was selected so as to minimize the effects both of recent overindulgence and of therapy. Since psychiatric diagnosis was less probable in the other three groups, a practical criterion included for members of all three was "without obvious mental or physical defect." In addition, AA members were required to have established a record of six months uninterrupted sobriety in that organization. Social drinkers were defined as women whose self-imputed use of alcoholic beverages ranged from "daily" to any frequency greater than once each month. In setting up the criterion for abstainers it was recognized that this group may include—besides absolute abstainers—persons who may on rare occasions use alcohol more as a courteous gesture than for any other reason. This criterion therefore included women whose use of alcohol ranged from "never" to a maximum of once per month.

A total of 132 subjects was tested. Following the discard of some records to make all four groups comparable as to age and length of education, 120 subjects made up the test group, including 24 active alcoholics, 34 AA members, 34 social drinkers, and 28 abstainers. Mean age for the entire group was 41.63 years and mean length of education was 13.16 years; applying F tests, comparability of groups in these two dimensions was established.

The test instrument consisted of copies of each of the two tests appended to a personal data sheet which asked information as to age, occupation, length of education, marital status, and drinking habits. AA members only were asked for details as to their drinking patterns and length of sobriety. In the case of tests given to the latter group, the Alcadd Test bore the notation "AA members are asked to answer this questionnaire on the basis of habits before entering AA."

Most subjects either lived in, or were hospitalized in, southwestern Connecticut; a few subjects were resident in New York City and a small number of active alcoholics were tested in a state alcoholism hospital in Hartford, Connecticut. Socioeconomically, the group tested appears to have been a middle-class one, there being no striking differences in this respect noted among the four subgroups. Length of continuous sobriety reported by AA members ranged from 6 to 176 months, with a mean length of 46 months and a median of 36 months.

Results

Recalling that higher scores indicate greater deviation in the cases of both tests used, the results shown in Table 1 appear to confirm the hypotheses tested in this research.

It will be noted that group mean scores on the Manson Evaluation make up a "ladder of scores" in the order: active alcoholics, AA members, social drinkers, abstainers. Of interest is the fact that scores for active alcoholics are much higher than those for AA members, the latter being closer to those for the social-drinker group. It would appear that, in terms of whatever dimensions are being measured by this test, AA members resemble social drinkers more than they resemble active alcohol addicts.

Results on the Alcadd Test also show such a "ladder" but with a reversal in that AA members obtained a higher mean score than did the active alcoholics. It will be recalled that AA members were asked to report their habits prior to their joining that group. No definite statement can be made—on present evidence—as to the reasons for this result; possible explanations which suggest them-

Table 1

Means and Standard Deviations of Scores for Four Test Groups, Using The Manson Evaluation and The Alcadd Test

		Manson Evaluation		Alcadd Test	
Test Group	*N*	Mean	*SD*	Mean	*SD*
Active alcoholics	24	38.6	12.2	25.5	11.9
AA members	34	23.4	12.3	41.4	9.7
Social drinkers	34	15.5	7.4	6.5	4.5
Abstainers	28	10.3	6.2	1.2	1.4

Table 2

The Manson Evaluation and The Alcadd Test: *t* Ratios of Differences Between Four Group Means

Comparison	Manson Evaluation	Alcadd Test
Active alcoholics vs. AA members	4.61	5.48
Active alcoholics vs. social drinkers	8.88	8.26
Active alcoholics vs. abstainers	10.48	10.57
AA members vs. social drinkers	3.16	18.57
AA members vs. abstainers	5.05	21.13
Social drinkers vs. abstainers	2.88	5.88

Note.—For degrees of freedom involved in these ratios, value of $t_{.01}$ ranged from 2.65 to 2.68.

selves include the following: (*a*) the higher AA scores may reflect a more accurate memory of aberrant habits than was possible for those closer in time to excessive drinking, (*b*) there might have been a tendency by the AA's to maximize difficulties considered as overcome, while the active addicts minimized recent patterns involving strong guilt feelings, and (*c*) the drinking patterns of the AA members in the test group may actually have been more deviant than those of the active alcoholics tested.

Statistical significance of the differences between group means on both tests is indicated by the *t* ratios given in Table 2.

It will be noted that all critical ratios are significant at well above the 1 per cent level. Discriminations tend to be more marked in results on the Alcadd Test, designed to discriminate among specific habit patterns, than those on the Manson Evaluation, which attempts measurement of less explicit personality dimensions. Despite the quantitative differences between results on the two tests, present findings appear to establish the validity of both, at least with respect to these samples.

Using Hoyt's approximation of the Kuder-Richardson formula, reliability coefficients were established at .92 in the case of the Manson Evaluation and at .97 in the instance of the Alcadd Test. These coefficients are almost identical to those established by the test author, which were .94 and .96, respectively. It would thus appear that, tested by the method of rational equivalence, both instruments are highly reliable.

Pearson product-moment correlation coefficients were derived from the scores of each group on both tests. Correlations were not significant except in the case of the active alcoholic group, for which the correlation amounted to + .50, significant at the 5 per cent level. It would appear that the two tests measure essentially dissimilar dimensions, the coincidence of which is explicit only in the case of the addicts tested.

A further inquiry into the results obtained concerned the predictive power of each test. In the present study, the effectiveness of these instruments is shown in Table 3.

Using the cutoff scores suggested by the test author, it is apparent that both tests are quite effective in predicting status, with the single exception that in this case the Manson Evaluation did not effectively identify alcoholics who are AA members. As noted above, the subjects in this group had a high average length of sobriety (46 months) and their mean test score on the Manson Evaluation was closer to the mean of the social drinkers than to that of the active alcoholics.

Discussion

Results of this research appear to confirm the validity and reliability of both tests used, as applied to female samples. Predictive powers established were in line with those claimed by the author, except that women AA members' status was not determined by use of the

Table 3

Predictive Powers of The Manson Evaluation and The Alcadd Test: Number and Percentage of Subjects in Four Test Groups Correctly Identified

Test Group	*N*	Manson Evaluation No.	Manson Evaluation %	Alcadd Test No.	Alcadd Test %
Active alcoholics	24	20	80	20	80
AA members	34	14	41	34	100
Social drinkers	34	31	91	32	94
Abstainers	28	27	96	28	100
Total	120	92	77	114	95

Manson Evaluation. Recalling that—in terms of the dimensions measured by that test—these AA members resembled the social drinkers tested rather than the active alcoholics, two possible implications are suggested: (*a*) that the Alcoholics Anonymous program of therapy has measurable effects, and (*b*) that alcoholics with certain personality configurations can benefit from the AA program while other alcoholics cannot do so.

It would appear that these tests might be of use in a number of practical applications, such as military or industrial screening, diagnostic differentiations, or aids to guidance counseling. At the same time, certain limitations of both instruments must be borne in mind. While the Manson Evaluation appears to be a valid and reliable predictor of a clinical syndrome, the etiological constituents of this syndrome are not presently identified. No evidence produced by this study overcomes the limitation stated by the test author in his original report: "It was not established that this test could differentiate alcoholics from nonalcoholic clinical deviants" (4, p. 204). As to the Alcadd Test, it is apparent that the content—dealing primarily with the subject's drinking habits—would limit its effective use in any situation which placed a premium on evasion.

It is suggested that further large-scale use of these tests, especially use involving other deviant groups, might yield data which, subjected to item and factor analyses, could point the way to a more effective understanding of alcohol addiction.

Summary

This report details results of research in which two questionnaires—the Manson Evaluation and the Alcadd Test—were used in testing 120 female subjects making up four test groups: active alcoholics, AA members, social drinkers, and abstainers. While the Manson Evaluation did not effectively identify the AA members in the test sample, all other results appear to confirm the clinical validity, reliability, and predictive power claimed by the author for both instruments.

Received September 20, 1955.

References

1. Barillas, M. G. Study of the validity and reliability of the Alcadd Test as a psychometric instrument for identifying male alcoholics. Unpublished master's thesis, Fordham Univer., 1950.
2. Manson, M. P. A psychometric differentiation of alcoholics from nonalcoholics. *Quart. J. Stud. Alcohol*, 1948, **9**, 175–206.
3. Manson, M. P. *The Manson Evaluation*. Beverly Hills, Calif.: Western Psychological Services, 1948.
4. Manson, M. P. A psychometric determination of alcoholic addiction. *Amer. J. Psychiat.*, 1949, **106**, 199–205.
5. Manson, M. P. *The Alcadd Test*. Beverly Hills, Calif.: Western Psychological Services, 1949.
6. Murphy, D. G. The validity and reliability of the Manson Evaluation and the Alcadd Test in the identification of female alcoholics. Unpublished master's thesis, Fordham Univer., 1955.
7. Sutherland, E. H., Schroeder, H. G., & Tordella, C. L. Personality traits and the alcoholic: a critique of existing studies. *Quart. J. Stud. Alcohol*, 1950, **11**, 547–561.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

A Comparison of the Bender-Gestalt Test and the Digit-Span Test as Measures of Recall[1]

Alexander Tolor[2]
USAF Hospital, Parks Air Force Base, California

This exploratory study was designed to investigate the ability of patients with organic brain disease and patients with psychogenic disorders to recall digits and Bender-Gestalt Test designs. Clinical experience has led some psychologists to conclude that difficulty in repeating digits forward or backward often reflects an organic defect (2, 10). However, the retention of digits also seems to be highly vulnerable to the effects of anxiety which may operate to cause impaired concentration and attention (7, 10). It would seem, therefore, that a disproportionately short memory span for digits could be ascribed to organic impairment or to impairment which is secondary to an emotional disturbance.

The Bender-Gestalt Test has only recently been used as a measure of memory for designs. The relationship between the number of Bender figures retained and other psychological variables has not been subjected to extensive investigation as yet. However, there are a few pertinent studies in the literature. Hanvik and Anderson (6) in comparing two groups of subjects having brain lesions with a group of control subjects found no significant differences between any of the groups as to number of figures recalled. Gobetz (4) assumed that, in the absence of an objective stimulus, emotional blocking would be greater for neurotic than for normal subjects and that the former would therefore recall fewer designs. This hypothesis failed to be supported by his data.

There are a number of differences between digit and Bender recall some of which are related to the nature of the task, the method of administration, and the type of response required. One of the most important differences pertains to the innocuous appearance of the Bender material and to the lesser social interaction between examiner and subject on the Bender test as compared to the Digit-Span test. The over-all effect is probably one of a more neutral emotional atmosphere during the presentation of Bender figures than during the presentation of digits.

Procedure

All patients were tested individually at the Neurological Institute of Columbia-Presbyterian Medical Center in accordance with the clinical procedures regularly employed by the staff of the Psychology Department.

With the exception of the vocabulary subtest, the whole Wechsler-Bellevue Scale, Form 1, was administered routinely. The Bender-Gestalt Test was presented as suggested by Bender (3). Following the copying of the figures, the designs were removed from sight and each patient was requested to reproduce from memory on a new sheet of paper as many of the designs as he could. When the patient seemed to have completed the task he was also asked to reproduce any parts of the designs that he could recall. Only one such additional urging was given.

Both whole and part figures were counted in computing the total number of Bender designs recalled. The figures which were reproduced from memory were also evaluated by the examiner for degree of distortion.

The Digit-Span subtest of the Wechsler-Bellevue Scale provided the data on total number of digits retained and on the number of digits recalled forward and backward. The difference between the Digit-Span weighted score and the patient's average Wechsler weighted score was also recorded to the nearest whole number.

All patients were tested prior to the determination of final diagnoses by members of the medical staff. The majority of the patients had been referred

---

[1] The opinions and conclusions expressed do not necessarily represent those of the Department of the Air Force.

[2] The data were collected while the author was staff psychologist at Neurological Institute, Columbia-Presbyterian Medical Center, New York City.

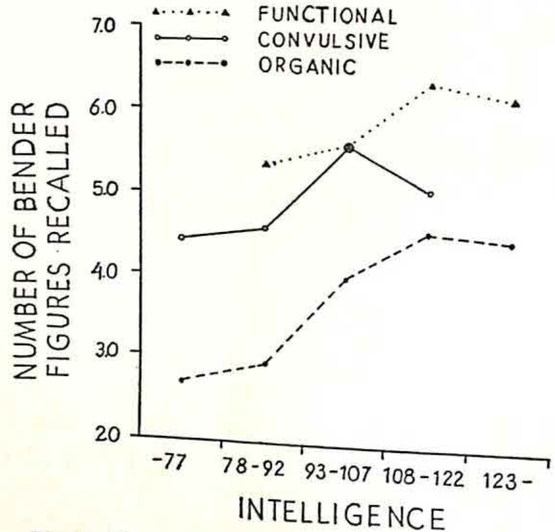


Fig. 1. Comparison of average number of Bender figures recalled by organic, functional, and convulsive patients at different levels of intelligence.

because their symptoms suggested the possibility of an organic brain disease. On the basis of their discharge diagnoses, the patients were separated into the following groups: an organic group, a seizure group, and a psychogenic group. The organic group contained 91 patients, all having some type of intracranial pathology. It included 23 patients with brain tumors, 16 with vascular pathologies, 4 with infectious diseases, 15 with head trauma, 27 with degenerative diseases, and 6 with hereditary-degenerative disorders. These classifications are very broad and arbitrary and were made only for the purpose of conveying some idea of the type of cases that were represented.

The psychogenic group contained 49 patients. Because of the unreliability of specific diagnostic categories no further nosological differentiation was attempted. However, it should be reported that only 3 patients were diagnosed as being psychotic, the remainder suffering from less severe personality disorders.

The 35 patients of the convulsive group were selected because their seizures, irrespective of whether they were petit mal, grand mal, or psychomotor, were of undetermined etiology. Thus, this population included only patients with idiopathic epilepsy.

Results

Table 1

Composition of the Three Groups

Item	Group Organic ($N=91$)	 Convulsive ($N=35$)	 Psychogenic ($N=49$)
Sex			
Male	57	14	23
Female	34	21	26
Age			
Range	12–72	12–56	12–64
Mean	44.1	27.6	35.4
SD	15.1	11.9	14.2
W-B Total IQ			
Range	54–129	61–121	79–136
Mean	90.8	94.2	110.8
SD	16.7	15.4	13.1

Table 1 contains information relative to the age, sex, and intelligence distribution of the three groups. There are obviously large age, sex, and intelligence differences among these groups. With respect to sex, only the proportion of males to females of the organic compared to the convulsive group differs significantly (.05 level). However, all the age differences are significant at the .01 level. In intelligence, the differences in mean IQ between the organic and psychogenic groups and between the convulsive and psychogenic groups are significant at the .01 level. The organic and convulsive groups do not differ significantly in intellectual functioning.

A significant and substantial correlation exists between the two measures of memory as is indicated by the coefficient of correlation of .43 ($\sigma_r = .06$) between digit recall and Bender recall for all groups combined.

Bender Recall

No significant departure from homogeneity was obtained by Cochran's test permitting the use of analysis-of-variance design. This technique shows that the mean number of Bender

Table 2

Analysis of Covariance

(Significance of differences among organic, psychogenic, and convulsive groups in Bender recall: effect of intelligence partialled out by analysis of covariance.)

Item	Source of Variance Total	 Within	 Between
Correlation	.50	.34	.91
df for *r*	173	171	1
b_{xy} value	.053	.033	—
Adjusted Sum of Squares	468.407	365.199	103.208
df	173	171	2
Adjusted Mean Squares	—	2.136	51.604

Note.—F for adjusted Mean Squares = 24.16; $p \leq .001$.

figures recalled in the three groups differs to a significant degree ($F = 44.38$; $p = < .001$). Since the two groups of patients with intracranial abnormalities scored consistently lower in intelligence than those without brain damage, it was necessary to control for the effects of intelligence in determining whether the recall of Bender figures is a function of clinical category. Table 2 shows that in adjusting for the effects of intelligence by analysis of covariance a high degree of significance is maintained.

Figure 1 compares the average number of Bender designs recalled at different levels of intelligence for the three groups. It will be noted that the organic patients consistently recall fewer Bender designs than the psychogenic patients and that the convulsive patients generally occupy an intermediate position irrespective of IQ range.

Table 3

Average Number of Bender Designs Recalled

Group	Uncorrected	Corrected
Organic	3.48	3.69
Convulsive	5.06	5.15
Psychogenic	5.98	5.53

The average number of Bender designs recalled in each of the three groups prior to and following adjustment for the effects of intelligence are shown in Table 3.

The relative difficulty in recall of the various Bender designs can be ascertained by inspection of Fig. 2. A chi square computed by Wilcoxon's (11) ranking method is 22.71, which yields a p of $< .004$, indicating that the difference between the relative frequencies with which the various designs are recalled is highly significant. Figure 2 also shows that there is a high degree of similarity among groups in the relative frequency of recall of specific Bender designs. For example, Configurations 4, 3, and A tend to be most difficult for subjects in all groups whereas Designs 8 and 6 tend to be easiest.

When the three groups are compared for the number of patients who produce severe Bender design distortions and minor distortions on recall, highly significant differences are found ($\chi^2 = 14.76$, $p = < .001$). Table 4 contains data indicating that the relative frequency of major distortions on recall is far greater in the organic group than in the other groups.

Table 4

Frequency of Bender Distortions in the Three Groups

Distortion	Organic ($N=88$)*	Convulsive ($N=35$)	Psychogenic ($N=49$)
Severe	33	3	7
Minor	55	32	42

* Three patients failed to recall any Bender figures and were therefore omitted in this analysis.

Digits Recalled

Since Cochran's test confirmed the assumption of equal variance, an analysis-of-variance technique could be applied to the data on the total number of digits recalled by the subjects in each of the three groups. The F of 20.17 yields a highly significant p of $< .001$. Table 5 illustrates that when the effects of intelligence are controlled for by analysis of covariance the adjusted means still differ significantly.

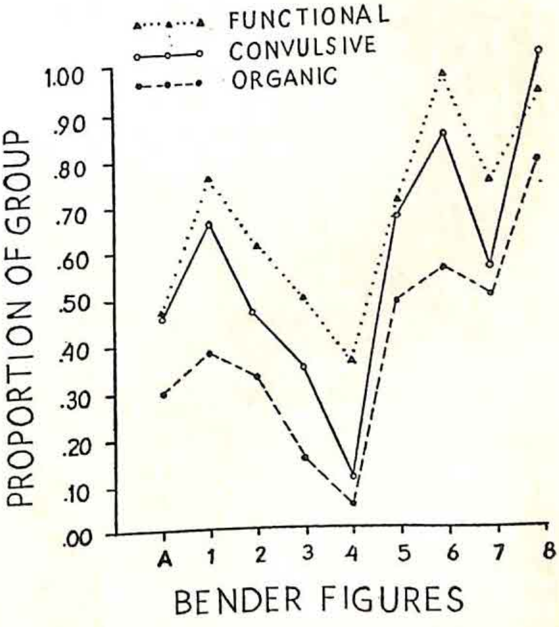


Fig. 2. Proportion of subjects in each of the groups recalling the different Bender designs.

Table 5

Analysis of Covariance

(Significance of differences among organic, convulsive, and psychogenic groups in total number of digits recalled: effect of intelligence partialled out by analysis of covariance.)

Item	Source of Variance		
	Total	Within	Between
Correlation	.70	.63	.94
df for *r*	173	171	1
b_{xy} value	.108	.101	—
Adjusted Sum of Squares	696.38	656.94	39.44
df	173	171	2
Adjusted Mean Squares	—	3.84	19.72

Note.—*F* for adjusted Mean Squares = 5.13; $p \leqq .01$.

When analysis-of-variance designs are applied to data on the recall of digits forward and to the recall of digits backward, the following values are obtained respectively: $F = 16.11$, $p \leqq .001$ and $F = 15.47$, $p \leqq .001$. In both sets of data when adjustments are made for effects of intelligence by analysis of covariance, significance drops to the .05 level. The average number of total digits recalled, and the number of digits retained forward and backward by the subjects in each of the three groups, are presented in Table 6.

The final comparison among groups is based on averages of differences between each patient's weighted Digit-Span subtest score and his Wechsler-Bellevue average weighted score. An analysis of variance for the three groups yields an *F* of 9.32, equivalent to $p < .01$. The means for the organic, convulsive, and psychogenic groups are − .69, − .09, and − .37, respectively. The data show that the convulsive patient tends to do approximately as well on the Digit-Span subtest as on the other Wechsler subtests. The psychogenic patient tends to have some degree of impairment in rote memory for digits whereas the organic patient demonstrates greatest impairment relative to his functioning in other areas. The difference between the organic and convulsive groups is significant at better than the .01 level, the difference between the organic and functional groups is significant at the .05 level, but the difference between the convulsive and functional groups is not statistically significant.

Table 6

Average Number of Digits Recalled

Digit Recall	Organic	Convulsive	Psychogenic
Total			
Uncorrected	9.43	10.97	12.20
Corrected	10.07	11.26	10.81
Forward			
Uncorrected	5.66	6.37	6.98
Corrected	5.94	6.49	6.37
Backward			
Uncorrected	3.77	4.60	5.22
Corrected	4.12	4.76	4.45

Discussion

The results of this study indicate that there is a higher correlation between number of digits recalled and intelligence ($r = .70$) than between the number of Bender figures recalled and intelligence ($r = .50$). When the digit-recall results are analyzed further, one finds a .60 correlation between intelligence and digits forward and a .65 correlation between intelligence and digits reversed. These correlations are somewhat higher than the *r* of .51 between intelligence and the Digit-Span test reported by Wechsler (10).

Although there is general agreement on the correlation between intelligence and number of digits recalled, there are some contradictory findings in the literature on the relationship between intelligence and Bender recall. Aaronson, Nelson, and Holt (1) obtained an *r* of .03 between MA on the Shipley-Hartford and Bender recall, and an *r* of − .10 between CQ and Bender recall. On the other hand, Peek and Olson (8) found a product-moment correlation of .34 between CQ and number of Bender designs recalled, and a correlation of .19 between the Shipley-Hartford and number of Bender figures retained. The present study would seem to indicate that a substantial relationship exists between intellectual functioning and Bender recall, one which is, however, less than the relationship between intellectual functioning and digit recall.

The data indicate also that the Bender recall is superior to digit recall as a measure of discrimination between organicity and non-organicity. However, with neither test are the differences among groups sufficiently large to permit prediction in individual cases.

The fact that the difference in recall between the organic and functional groups is greater on the Bender test than on the Digit-Span test seems to support the impression that the Bender is less susceptible to the effects of anxiety than the Digit-Span test.

It is interesting that the results obtained on the relative difficulty in recall of various Bender designs differ somewhat from those reported by Goodstein *et al.* (5). In both studies Designs 3 and 4 were found most difficult, but the ease with which Figures 8 and 6 were recalled by the patients in all of our groups was at variance with the results of the other study. Furthermore, the subjects used in the present study did not find the first three designs to be particularly easy. On the contrary, considerable difficulty was encountered in reproducing Design A. These differences could be attributed to differences in subjects used or to differences in experimental methodology. Inasmuch as there is some evidence (9) that the various designs differ in symbolic significance, it might be useful to undertake an investigation of the relationship between design difficulty and the emotional value which the designs have for subjects.

Summary

Patients having organic brain disease, idiopathetic seizures, and psychogenic disorders were compared for their performance on the Digit-Span test and on the Bender-Gestalt Test. The results indicated that there are significant differences among groups on both measures of immediate recall. These differences hold even when adjustments are made for the effects of intelligence. In Bender recall the psychogenic group does best, the organic group does least well, whereas the convulsive group occupies an intermediate position. On the Digit-Span test the order of recall is convulsive group, psychogenic group, and organic group, with the latter retaining the fewest digits. The differences among groups are more pronounced on the Bender test than on the Digit-Span test. The level of difficulty of individual Bender designs was also determined. Some of the results were discussed in terms of the differential susceptibility of the measures of recall to the effects of anxiety.

Received September 20, 1955.

References

1. Aaronson, B. S., Nelson, S. E., & Holt, Shirley. On a relation between Bender-Gestalt recall and Shipley-Hartford scores. *J. clin. Psychol.*, 1953, **9**, 88.
2. Allen, R. M. The test performance of the brain injured. *J. clin. Psychol.*, 1947, **3**, 225–230.
3. Bender, Lauretta. A visual motor gestalt test and its clinical use. *Amer. Orthopsychiat. Ass. Res. Monogr.*, 1938, No. 3.
4. Gobetz, W. A quantification of the Bender-Gestalt Test on normal and neurotic adults. *Psychol. Monogr.*, 1953, **67**, No. 6 (Whole No. 356).
5. Goodstein, L. D., Spielberger, C. D., Williams, J. E., & Dahlstrom, W. G. The effects of serial position and design difficulty on recall of the Bender-Gestalt Test designs. *J. consult. Psychol.*, 1955, **19**, 230–234.
6. Hanvik, L. J., & Anderson, A. L. The effect of focal brain lesions on recall and on the production of rotations in the Bender-Gestalt test. *J. consult. Psychol.*, 1950, **14**, 197–198.
7. Moldawsky, S., & Moldawsky, Patricia C. Digit span as an anxiety indicator. *J. consult. Psychol.*, 1952, **16**, 115–118.
8. Peek, R. M., & Olson, G. W. The Bender-Gestalt recall as an index of intellectual functioning. *J. clin. Psychol.*, 1955, **11**, 185–188.
9. Suczek, R. F., & Klopfer, W. G. Interpretation of the Bender Gestalt Test: the associative value of the figures. *Amer. J. Orthopsychiat.*, 1952, **22**, 62–75.
10. Wechsler, D. *The measurement of adult intelligence.* (3rd Ed.) Baltimore: Williams & Wilkins, 1944.
11. Wilcoxon, F. *Some rapid approximate statistical procedures.* Stamford, Conn.: American Cyanamid Co., 1949.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

The Relation of the MMPI to the Edwards Personal Preference Schedule on a College Counseling Center Sample[1]

Reed M. Merrill
University of Utah

and Louise B. Heathers
University of Washington

The Minnesota Multiphasic Personality Inventory (MMPI) is a well-known test instrument which is frequently used in counseling college students even though its normative data are based on a general adult population. Most of the scales of this test were constructed empirically by selecting items which differentiated a diagnosed adult clinical sample from a normal adult sample (10). The purpose of the research to be reported here is to compare the MMPI with a new personality test standardized on a college population and designed to measure the relative strength of 15 personality needs or dimensions.

The new test is the Edwards Personal Preference Schedule (PPS). It consists of 210 forced-choice items. Each pair of items is matched approximately for mean social desirability value, as previously judged by college students, to minimize the effect of social desirability on item choice. Items measuring one need are paired twice with each of the remaining 14 needs; hence the maximum raw score on any need is 28. Since the items are paired, the total raw score on the test will be the same for all persons. Thus the PPS reflects only the relative strength of competing needs and attitudes rather than the absolute strength of any need. Hence this test clearly has the advantages as well as the disadvantages of an ipsative scale.

The needs measured by the PPS are those indicated in the manual (4) which gives Edwards' description of the variables. The PPS sets out to measure the needs defined by Murray (14) and uses similar nomenclature in titling the needs. Test items were originally chosen for the scales on an a priori basis in consultation with other psychologists.

To obtain an estimate of the degree of relationship present between these two instruments tetrachoric correlations were computed. The sample of subjects was dichotomized as close to the median score of the group as possible. The most diverse of the dichotomies was one of 44 per cent of the sample vs. 56 per cent of the sample. Gohsen's method (5) of estimating tetrachoric *r*s was used. Since tetrachoric *r*s are less reliable than Pearson *r*s, Pearson *r*s were computed when the relationship between two variables as reflected by the r_t appeared very highly significant. The Pearson *r*s were somewhat lower than the r_ts but the level of significance was not changed. Correlations were computed not only for the standard validating and clinical scales of the MMPI, but also for ten experimental scales derived from the MMPI. These were Gough's Dominance, Responsibility, and Status scales (7, 8, 9), Drake's Social-Introversion (3), Navran's Dependence (15), Taylor's Manifest Anxiety (16), Winne's Neuroticism (17), Cook and Medley's Pharasaic-Virtue and Hostility scales (2), and an unpublished Social Desirability scale constructed from MMPI items by Edwards (4).

The sample consisted of all available presumably valid records of males seen at a university counseling center for vocational or educational advisement who had taken both the MMPI and PPS during the same counseling contact. Following Hathaway and Meehl's suggestion (11), records with $L \geqq 10$ or $F \geqq 16$, or $Q > 30$ were eliminated (4 *S*s). In addition, since $F - K$ scores of 21 or more are most frequent in selection situations where persons are deliberately attempting to make good impressions (12), records where $F - K$ was $\geqq 21$ were eliminated (6 *S*s). One other criterion for inclusion in the sample was used. On the PPS there are 15 item pairs which are exact duplicates; the number of agreements on these 15 items is used as a measure of the consistency of response. Edwards suggests that with a consistency score of 8 or less the scale scores are highly suspect. Six per cent of his male normative sample obtained consistency scores of 8 or less. To find a similar cutting point on the MMPI, a 10 per cent random sample of 1,500 male MMPI records were scored on the 16 duplicate items of the MMPI. For these 150 records 5 per cent of the *S*s obtained consistency scores of 12 or less. Hence records were eliminated where the PPS consistency score was 8 or less or where the derived MMPI consistency score was 12 or less (12 *S*s). Of the original 177 available records, 155 remained. These included veterans and nonveterans, college (82.6 per cent) and noncollege clients. The most typical client was a single, 23-year-old college sophomore.

---

[1] Based, in part, on a paper presented at the convention of the American Psychological Association, San Francisco, California, 1955.

Table 1 shows the means of the sample in the standard T scores of the usual MMPI scales and of the PPS. T score means are not shown for the experimental MMPI scales since such scores are not available on all of these scales and since, where they are available, they are based on quite different populations. Table 1 shows that persons who come to a counseling center for educational or vocational counseling differ considerably from the respective normative samples, which is to be expected. Also, as one would expect, deviations from the normative group are more marked on the MMPI where the normative population is a general adult sample than on the PPS where the normative population is a college sample. This finding is consistent with other studies (6) of the MMPI which have indicated that college students in general differ from the MMPI normative group. On the PPS the counseling client deviates most markedly from the normative college males on three scales: they rate themselves especially low on the heterosexuality scale, especially high on the achievement and endurance scales. These differences are interest-

Table 1

Client Score Distributions on the PPS and MMPI ($N = 155$)

PPS				MMPI				Experimental MMPI		
Scale	Mean	σ	T*	Scale	Mean	σ	T*	Scale	Mean	σ
Ach	16.95	4.05	53	L	3.34	1.95	50	Dom	16.87	3.55
Def	11.92	3.36	52	F	4.70	2.97	55	Resp	21.21	3.53
Ord	11.23	4.47	52	K	15.05	4.52	55	Status	22.58	3.59
Exh	13.60	3.78	49	$F-K$	−10.38	6.30	—	Si	27.04	10.14
Aut	15.01	4.10	51	Hs	4.35	3.64	49	Dep	20.29	10.05
Aff	14.13	4.24	48	$Hs+K$	12.12	3.47	52	Man. Anx.	15.73	7.98
Int	16.22	4.99	50	D	20.22	5.18	58	Neur	5.50	4.01
Suc	10.27	4.50	48	Hy	20.48	4.58	56	Pv	16.62	7.68
Dom	17.05	5.04	49	Pd	17.02	5.13	58	Ho	17.62	7.68
Aba	13.11	5.34	52	$Pd+K$	23.06	4.94	60	Soc. Des.	30.79	5.63
Nur	13.36	4.68	48	Mf	27.08	5.20	63			
Chg	15.88	4.37	51	Pa	9.27	2.89	53			
End	13.74	5.77	53	Pt	13.48	8.10	54			
Het	15.45	5.84	45	$Pt+K$	28.58	5.94	62			
Agg	12.09	4.84	48	Sc	12.34	7.88	53			
				$Sc+K$	27.46	6.38	59			
				Ma	17.15	3.88	57			
				$Ma+K$	20.14	3.79	58			

* Ts are approximations.

ing and are developed in other research (13) but here our concern is with a comparison of the two tests.

Table 2 presents the tetrachoric correlations between the MMPI and the PPS scales. Most of the correlations found are not statistically significant. However, more significant correlations were obtained than one would expect by chance alone. In addition, using the technique suggested by Brožek and Tiede (1), the probabilities of the compound event of obtaining n significant critical ratios in a series of N tests of significance was computed. The critical ratio obtained was 16.33 so that the number of significant relationships found is far beyond chance expectations.

Where significant correlations exist, the direction of the relationship makes sense with clinical expectations developed from MMPI research. For example, Change correlates positively with *Ma;* Aggression correlates negatively with *L* and *K*, positively with non-*K*-corrected *Pd;* Abasement correlates positively with *F*, negatively with *K*, positively with non-*K*-corrected *Pt* and *Sc*. Three PPS scales are of particular interest in relation to the usual MMPI clinical scales. Abasement is most clearly related to the clinical scales. On the other hand, Deference and Dominance are inversely related to the clinical scales. Clients who rate their guilt and inadequacy feelings relatively high on the PPS are most likely to be abnormal as measured by the MMPI. Clients who have accepted either a dominant, strong role or a more conforming, follower role in interpersonal relationships on the PPS tend to be least like deviants as measured by the MMPI. Most of the PPS scales, however, show little relationship with the MMPI clinical scales.

The effect of the *K* correction of the MMPI on the correlations is of interest. College counselors have sometimes felt that *K*-corrected scores underestimated the adjustment level of college students since, in general, they tend to obtain high *K* scores. To a degree this objection probably still holds when a correction developed from abnormal groups is applied to the responses of fairly normal groups; nonetheless, the *K*-corrected scores appear of value in using the MMPI with the latter groups since the *K* scale does tend to correct for test-taking attitudes.

On the PPS the social desirability stereotype of the items is partially controlled so that it is less easy to respond only in terms of culturally approved responses. The *K* correction tends to have the same effect on the MMPI scales as indicated by the correlational data of this study. Of the nine r_ts significant at the 5 per cent level when MMPI scales are not *K* corrected, only one remains significant at the 5 per cent level when the *K* correction is applied; of the nine r_ts significant at the 1 per cent level, two remain significant at the 1 per cent level when the *K* correction is applied, three drop to the 5 per cent level of significance, and the remaining four become nonsignificant. Thus a partial control of the biasing effects of the cultural stereotype of normal behavior on the MMPI clinical scales reduces the relationship between the MMPI and the PPS.

The results of the correlation of the PPS with the MMPI experimental scales are quite striking. The significant correlations are again generally consistent with clinical expectations but relatively low. The two Dominance and Hostility scales correlate positively but to a lesser degree than one might expect. Responsibility correlates negatively with Aggression, but the expected negative correlations with Autonomy and Change are too low to be significant. Attitudes characteristic of persons scoring high on the Status scale are directly related to Dominance and inversely related to Succorance and Abasement. These findings are consistent with our usual cultural patterns.

Of general interest is the consistency of the direction of the correlations of the MMPI experimental scales with the PPS if the Social Desirability, Status, Responsibility, and Dominance scales are viewed as measuring socially approved or "healthy" behavior and the remaining scales are viewed as measuring socially disapproved or "sick" behavior. Of the PPS scales, Dominance, which is probably most closely related to a normal MMPI profile, has positive relationships with the "healthy" MMPI experimental scales and negative correlations with the "sick" scales. Of the PPS scales, Abasement, which is most closely related to an abnormal MMPI profile, has the opposite relationship—negative relationships with the "healthy" scales and positive relationships with the "sick" scales. Exhibition and Aggression show similar inverse patterns where the correlations reach a significant level. It is noteworthy that the attempt by the PPS to rule out the social desirability preference value of an item is not entirely successful; four of the correlations between the Social Desirability scale based on the MMPI and PPS variables are significantly above zero.

The pool of *S*s was too small to obtain statistically reliable differences on MMPI scales between *S*s with varying profiles on the PPS. However, some observations on such profiles appear pertinent. A pattern that occurs rather frequently is one with Deference and Endurance high and Autonomy and Aggression low or vice versa. Those with the high Aggression-Autonomy pattern tend to give a much more deviant MMPI profile than those with a high Deference-Endurance profile. The Aggression-Autonomy group tend to be lower on the "healthy" experimental scales and higher on the "sick" experimental scales. Those with high Succorance scores are quite similar to those with high Abasement scores. It appeared that a high Succorance plus a high Heterosexuality score or, to a less extent, a high Abasement score plus a high Heterosexuality score, identified those in this group who scored especially high on the Dependence scale. Again the data suggest that a high Heterosexuality score on the PPS is related to lack of adjustment rather than to adjustment. There seems to be a need to deny conforming, socially acceptable attitudes rather than a lack of anxiety about sexual impulses.

In summary, the correlations between the scales of the PPS and the MMPI on a college counseling center sample are not high. The relative ratings of a client's needs as he views them on the PPS may or may not indicate the presence of personality symptom formation as defined by the MMPI. This finding is not surprising in view of the fact that the PPS is an ipsative scale which does not reflect the absolute level of need and the fact that psychiatric diagnoses, crucial to the construction of the MMPI, are often related more to the manner of satisfying need than to the nature of the needs themselves. However, where relationships are found they tend to be consistent with MMPI research data.

Both tests will probably be found useful in college counseling centers. Just as the Allport Study of Values contributes something unique when used in combination with the Strong Vocational Interest Blank, so does the PPS make a contribution distinct from that of the MMPI. The PPS shows the relative weight a person gives to various personal needs and the MMPI the degree of response similarity to well-defined clinical groups.

Received October 17, 1955.

References

1. Brožek, J., & Tiede, K. Reliable and questionable significance in a series of statistical tests. *Psychol. Bull.*, 1952, **49**, 339–341.
2. Cook, W. W., & Medley, D. M. Proposed hostility and pharisaic-virtue scales for the MMPI. *J. appl. Psychol.*, 1954, **38**, 414–418.
3. Drake, L. E. A social I. E. scale for the MMPI. *J. appl. Psychol.*, 1946, **30**, 51–54.
4. Edwards, A. L. *Manual for the Edwards Personal Preference Schedule.* New York: Psychological Corp., 1954.
5. Edwards, A. L. *Statistical methods for the behavioral sciences.* New York: Rinehart, 1954.
6. Goodstein, L. D. Regional differences in MMPI responses among male college students. *J. consult. Psychol.*, 1954, **18**, 437–441.
7. Gough, H. G. A new dimension of status: I. Development of a personality scale. *Amer. sociol. Rev.*, 1948, **13**, 401–409.
8. Gough, H. G., McClosky, H., & Meehl, P. E. A personality scale for dominance. *J. abnorm. soc. Psychol.*, 1951, **46**, 360–366.
9. Gough, H. G., McClosky, H., & Meehl, P. E. A personality scale for social responsibility. *J. abnorm. soc. Psychol.*, 1952, **47**, 73–80.
10. Hathaway, S. R., & McKinley, J. C. *Minnesota Multiphasic Personality Inventory; manual.* (Rev. Ed.) New York: Psychological Corp., 1943, 1951.
11. Hathaway, S. R., & Meehl, P. E. *Atlas for the clinical use of the MMPI.* Minneapolis: Univer. of Minnesota Press, 1951.
12. MacLean, A. G., Tait, A. T., & Catterall, C. D. The F minus K index on the MMPI. *J. appl. Psychol.*, 1953, **37**, 315–316.
13. Merrill, R. M., & Heathers, Louise B. A comparison of guidance center clients with the normative populations of the Edwards Personal Preference Schedule. Unpublished manuscript.
14. Murray, H. A., *et al. Explorations in personality.* New York: Oxford Univer. Press, 1938.
15. Navran, L. A rationally derived MMPI scale to measure dependence. *J. consult. Psychol.*, 1954, **18**, 192.
16. Taylor, Janet A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

The Influence of Social Desirability on Discrepancy Measures Between Real Self and Ideal Self[1]

Douglas T. Kenny
University of British Columbia

A potentially fruitful and commonly accepted clinical hypothesis is the formulation that a discrepancy between an individual's conception of his real self and ideal self is indicative of maladjustment and that as psychotherapy proceeds the discrepancy between the two selves should decrease. The present investigation was undertaken to highlight the necessity for controlling an important variable in the valid testing of this hypothesis, namely, that of social desirability.

In obtaining a discrepancy index between real self and ideal self two similar methodological procedures have been commonly employed. One method (2) consists in having an individual rate a number of dynamic personality traits on, say, a 7-point rating scale under two instructional conditions. The person is asked first to rate his conception of his real self and secondly to rate his notion of his ideal self. The discrepancy measure is provided by the absolute sum of the differences between the pairs of trait items on the real-self and ideal-self ratings. Another procedure for obtaining an operational index of discrepancy has been through Q technique (7). In this method, the individual sorts a given group of statements or traits into successive intervals that follow a quasi-normal frequency distribution. That is, the person is instructed to place a specified number of statements into the different categories on a continuum from most descriptive to least descriptive. The individual performs two Q sorts, a real-self sort and an ideal-self sort. The correlation between these two Q sorts provides the index of discrepancy. The lower the correlation is, presumably, the greater the maladjustment.

Edwards and Horst (4) have hypothesized, on the basis of Edwards' finding that social desirability correlates highly with the probability of endorsement of personality questionnaire items, that the social desirability factor should operate in Q sorts. That is, statements which are socially desirable should be judged by individuals as most characteristic of themselves. By an extension of their hypothesis, it may be postulated that social desirability would enter also as an obtrusive variable into the rating scale procedure to real-ideal self discrepancies. In fact, it is not unreasonable to suppose that some of the discrepant findings between various studies in this area may have occurred as a partial or direct result of the failure to control for this variable. For instance, while the studies of Bills (1) and Roberts (6) have offered some positive support for the hypothesis that a discrepancy between real self and ideal self is indicative of personality disturbance, Zimmer's (8) investigation yielded negative results. However, the latter study would be open to criticism if it were shown that the personality traits employed were "loaded" with the social desirability variable. It is one of the purposes of this study to show that such was the case.

A secondary purpose of the present investigation is to determine whether or not social desirability enters to the same extent into personality questionnaire items, rating scale items, and items in Q sorts. There was the initial expectancy that social desirability would play a differential role in the three personality-evaluation techniques. The rationale for this expectancy was simply that individuals would censor their self-descriptions to a greater extent in the more highly structured techniques than in the less structured ones. The order of structure, from most to least, would seem to be from the questionnaire items, to the rating scales, and then to the Q sorts.

---

[1] This article represents a modification of a paper presented at the American Psychological Association, San Francisco, 1955.

Method

Social desirability ratings. Twenty-five self concepts, used by Zimmer (8), were scaled for their social desirability on a 7-point scale by 67 undergraduate university students (28 females and 39 males). Zimmer's criteria for these traits were that they must (*a*) occur between six to ten times in one million words, as judged by the Thorndike-Lorge word count; (*b*) reflect dynamic content; and (*c*) not represent "a cultural stereotype of a desirable or undesirable trait" (8, p. 447). No explicit reference was made as to how the latter two criteria were assessed.

The instructions to the judges were to rate each trait on a continuum of social desirability. The median rating for each trait served as its social-desirability scale value. When these traits are ordered from most socially desirable to least socially desirable, the following rank order is obtained: respectful, energetic, orderly, refined, ambitious, trusting, spontaneous, precise, determined, obedient, economical, ardent, persistent, deliberate, cautious, conventional, leisurely, daring, sentimental, poetic, emotional, wary, dominant, meek, and lusty.

Questionnaire endorsement. Each of the 25 trait terms were written in the form of a positive assertion, such as, "I am respectful," "I am energetic," and so on. Another independent sample of 65 university students, composed of 31 females and 34 males, answered these personality questionnaire items as to whether or not they were characteristic of themselves. As in Edwards' study (3), the proportion of subjects stating that each trait was characteristic of themselves was computed, and these proportions served as the self-endorsement values for the respective traits.

Rating scale descriptions. A third independent sample of 58 university students (28 females and 30 males) rated themselves on the 25 traits, once for real self and once for ideal self. As in Zimmer's study (8, p. 447), the real-self scale went from "*7*. I am a very . . . person" to "*1*. I am definitely not a . . . person," and the ideal-self scale went from "*7*. I would like to be a very . . . person" to "*1*. I definitely don't want to be a . . . person." For the respective measures of real self and ideal self, the median rating scale value for each item served as its probability of endorsement value.

Table 1

Social-Desirability Correlates on Questionnaire and Rating Scales

Variable	Questionnaire	Real Self	Ideal Self
Social Desirability	.82**	.81**	.82**
Real Self			.63**

** $p < .01$.

Q sorts. A fourth and small independent sample of eight university students, equally divided between the sexes, performed three Q sorts on the traits. The traits were first sorted for real self, then ideal self, and thirdly for how the subjects thought the traits arranged themselves in terms of social desirability. The following quasi-normal frequency distribution was used: From *0* (least characteristic) to *6* (most characteristic) the seven steps required frequencies of 2, 3, 4, 7, 4, 3, and 2, respectively.

While it is usually considered desirable to have a larger number of categories and a larger sample of trait statements for Q sorts, the main interest was to demonstrate the possible operation of social desirability on Q sorts and not in conducting a Q study per se.

Results

In order to show the relationship between social desirability and the probability of endorsement on the questionnaire and rating scales, correlations were determined between the rank order of the traits on the social-desirability continuum and those based on probability of endorsement. The resulting rank-order correlations are shown in Table 1. The correlation of .82 between desirability and probability of endorsement confirms Edwards' (3) original finding of a correlation of .87 between probability of endorsement of 140 personality items based on Murray's needs and social-desirability scale values. Further, Hanley (5) has obtained similar results, reporting correlations of .89 and .81 between desirability rating and probability of endorsement of a random sample of items from the *Sc* and *D* scales of the MMPI, respectively.

Table 1 indicates also that the social-desirability variable intrudes into the real-self and ideal-self rating scales to a similar magnitude as it does into questionnaire items. The correlation of .63 between the rank orders of the real self and ideal self is undoubtedly a partial reflection of social desirability. Thus, the results indicate no support for the hypothesis that social desirability plays a differential role in these two personality-evaluation techniques, at least for the items employed in this study.

The relationships existing between the various Q sorts are given in Table 2. The values presented under the real self vs. social desirability columns represent the correlations between each individual's real-self sort and his social-desirability sort and between each individual's real-self sort and the redistribution of the 67 judges' median social-desirability values of the traits to fit the forced intervals used in the Q sorts. A similar analysis, for ideal self, is expressed in the last two columns of Table 2. The correlations between social desirability and the real-self and ideal-self sorts are highlighted in this table. Thus the data support the Edwards and Horst (4) formulation that social desirability should enter in Q sorts. It is of some interest to note that the real self and ideal self correlate significantly with the independent group conception of social desirability in every case but two.

Table 2

Individual Q-Sort Relationships

Subject Number	Real Self vs. Ideal Self	Real Self vs. Social Desirability		Ideal Self vs. Social Desirability	
		Own	Group	Own	Group
1	.73**	.49**	.41**	.16	.46**
2	.31	.01	.47**	.01	.47**
3	.78**	.65**	.47**	.86**	.82**
4	.75**	.30	.37*	.27	.25
5	.76**	.60**	.48**	.58**	.68**
6	.52**	.29	.26	.31	.52**
7	.59**	.72**	.65**	.83**	.69**
8	.85**	.64**	.33*	.71**	.38*

* $p < .05$.
** $p < .01$.

Table 3

Composite Q-Sort Relationships

	Real Self	Ideal Self
Social Desirability	.66**	.59**
Real Self		.82**

** $p < .01$.

To assess the influence of social desirability on the Q sorts taken as a whole, the eight scores for each trait item were averaged to form an averaged score value. On the basis of their averaged score values, the 25 traits were redistributed into the categories of the quasi-normal frequency distribution. This procedure was followed separately for the real-self, ideal-self and social-desirability sorts. Table 3 reports the intercorrelations between these averaged self-assessments. Again the operation of social desirability on Q sorts is confirmed.

Discussion

A major implication of the present findings is that social desirability operates in a significant fashion in all three personality-evaluation techniques. No evidence was found to support the initial expectancy that social desirability would play a differential role in the three procedures.

The results also have practical implications for investigators who test the twofold clinical hypothesis that maladjustment is indicated by a discrepancy between real self and ideal self and that the correlation between the two selves should decrease as therapy proceeds. Unless the social-desirability variable is controlled, the specific variance in the difference score between real self and ideal self will be negligible because social desirability will cancel out any reliable difference between the two selves. That is, the "inflated" correlation between real self and ideal self has an important influence on the specific and error variance measured by any computed discrepancy (difference) score. The net effect of a high correlation between the two selves is to reduce specific variance and to increase error variance in the discrepancy score. Hence, it would be fruitless to investigate whether or not a discrepancy index is related to personality maladjustment if social desirability produced a high correlation between real self and ideal self. Furthermore, the error variance in the difference score will be large if the respective reliabilities of the real-self and ideal-self measures are low.

It is obvious that the specific variance of the discrepancy score may be increased by having a large number of traits rated and by controlling the influence of the social-desirability variable. The former procedure has, of course, its exceptions. But, how might one control for the latter variable? For structured Q samples, Edwards and Horst (4) suggest that it can be controlled by having for each of the two or more kinds of items a balancing in terms of social-desirability scale values. For instance, if the structured sample is composed of only surgency and desurgency items, they would have for every item in the surgency group one in the desurgency group matched for social desirability. However, they make no suggestions for unstructured Q samples or for rating scale items, such as used in this study. For unstructured samples, social desirability could be partially controlled by utilizing traits which are ambiguous in terms of social desirability. Ambiguity could be defined in terms of the standard deviations of the social-desirability ratings. Another procedure for controlling social desirability in unstructured samples would be to use only neutral trait items, that is, those items near the median of judged desirability.

Finally, it should be noted that the present data cast doubt upon Zimmer's negative findings on testing the hypothesis that a real-ideal discrepancy indicates conflict. It will be recalled that the present study employed the same traits as those used by Zimmer. Because the present findings indicate that these traits are "loaded" with social desirability, stereotypic responses to them undoubtedly occurred in Zimmer's study. As a result, little personal involvement may have entered into his subjects' judgments. In short, unequivocal evidence against the discrepancy hypothesis requires rigorous control over the social-desirability variable.

Summary

This study was designed to indicate the necessity for controlling the social-desirability factor in testing the twofold clinical hypothesis regarding discrepancy between real self and ideal self and to assess the relative influence of social desirability on personality questionnaires, rating scales, and Q sorts. The findings support the conclusion that social desirability enters into the three personality techniques to about the same extent. The results further suggest that Zimmer's conclusion that discrepancies between real self and ideal self are not indicative of conflict is open to question because of his failure to control for the social-desirability factor.

It was emphasized that the control of social desirability is an indispensable aspect of any study testing the clinical hypothesis about real-ideal self discrepancy. Procedures for increasing the specific variance of discrepancy indices and for controlling the social-desirability variable were discussed.

Received October 25, 1955.

References

1. Bills, R. E. A validation of changes in scores on the index of adjustment and values as measures of changes in emotionality. *J. consult. Psychol.*, 1953, **17**, 135–138.
2. Bills, R. E., Vance, E. L., & McLean, O. S. An index of adjustment and values. *J. consult. Psychol.*, 1951, **15**, 257–261.
3. Edwards, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *J. appl. Psychol.*, 1953, **37**, 90–93.
4. Edwards, A. L., & Horst, P. Social desirability as a variable in Q-technique studies. *Educ. psychol. Measmt*, 1953, **13**, 620–625.
5. Hanley, C. Judged social desirability and probability of endorsement of items on the MMPI *Sc* and *D* scales. *Amer. Psychologist*, 1955, **10**, 404–405. (Abstract)
6. Roberts, G. E. A study of the validity of the index of adjustment and values. *J. consult. Psychol.*, 1952, **16**, 302–304.
7. Rogers, C. R., & Dymond, Rosalind F. (Eds.) *Psychotherapy and personality change.* Chicago: Univer. of Chicago Press, 1954.
8. Zimmer, H. Self-acceptance and its relation to conflict. *J. consult. Psychol.*, 1954, **18**, 447–449.

Journal of Consulting Psychology
Vol. 20, No. 4, 1956

The Evaluation of the Significance of Differences Between Scaled Scores on the WAIS: The Perpetuation of a Fallacy

H. Gwynne Jones

Institute of Psychiatry (Maudsley Hospital), University of London

In the WAIS manual, Wechsler claims that ". . . about two-thirds of the observed differences between scores on any two tests do not exceed three scaled score points. Differences as large as five points may be unusual enough to be noteworthy; smaller differences should be recognized as more common, therefore less likely to be significant" (3, p. 18). In making this statement Wechsler is inviting the clinical psychologist using the test to conclude that such discrepancies, being rare, are indicative of pathology, a practice already common, largely as a result of Wechsler's earlier writings (2). In fact, differences of this order of magnitude are very common, even in the standardization sample, as is evident from Wechsler's own argument.

There are 55 possible pairings for the eleven tests in the scale. Wechsler examined the corresponding 55 distributions of differences for the 25–34 age group, found that the median value of the standard deviations was approximately three, and deduced the proposition quoted above. His choice of a difference of five points as a critical level represents an *SD* value of approximately 1.7 and, therefore, significance at the 9 per cent level. This percentage does not, however, refer to the proportion of individual subjects displaying such a discrepancy but to the proportion of differences of which 55 occur for each subject. Therefore, on Wechsler's reasoning, some four to five discrepancies of this magnitude are to be expected in any one record.

Wechsler's statement is, of course, more apt if one is concerned, not with a difference between *any* two tests, but with a difference between two specified tests as such specific pairings occur only once in each record. Such usage, however, is only justified when there are prior grounds for selecting the specific difference examined. Thus, for example, if it is predicted on independent grounds that a certain subject will achieve an abnormally low score on Block Designs as compared to Vocabulary then Wechsler's method may be applied though, of course, the direction of the difference is then of importance and a one-tailed test may be applied. In such circumstances, however, it would be unwise to make use of Wechsler's approximate average standard deviation as, owing to the wide range of correlation between different tests, the magnitude of the standard deviation varies considerably. Thus, for example, the correlation of .81 between Vocabulary and Information yields a standard deviation of differences below two scaled score points while the correlation of .30 between Object Assembly and Digit Span raises the value to over 3.5.

As the writer has only recently received the WAIS test material, empirical data derived from that test cannot be quoted in support of the arguments advanced in this paper. Similar reasoning can, however, be applied to test results derived from the earlier Wechsler-Bellevue Scale (2). The scaled scores of 32 Kansas Highway Patrolmen, rated as well adjusted, are reported by Rapaport (1, p. 521). If examined for subtest discrepancies, these data show at least one difference of at least five scaled score points for all but one subject. In all 32 records there are 251 differences of this magnitude, representing some 14 per cent of the 1,760 differences involved. Sixty-one differences, or 3.5 per cent of the total, are of seven or more scaled score points, and eight differences reach ten points or more. When the 110 possible specific, directional discrepancies are examined, 56 of these fail to reach six scaled score points in any record. Of the 54 which achieve this magnitude or greater, 17 occur once, 19 twice, 7 three times, 7 four times, and 4 five times.

Wechsler promises to discuss fully the relation of his findings concerning subtest differences to problems of patterning in a forthcoming book. It is to be hoped that he will clarify the issues discussed in this paper. If the wide range of random scatter found in normal protocols were fully realized much of the effort now expended in research on scatter analysis would be more fruitfully applied. Though group patterns may be established, the range of individual deviations will be so great as to render the procedure useless for individual application by clinical psychologists.

Received September 19, 1955.

References

1. Rapaport, D. *Diagnostic psychological testing*, Vol. 1. Chicago: Year Book Publishers, 1945.
2. Wechsler, D. *The measurement of adult intelligence.* (3rd Ed.) Baltimore: Williams & Wilkins, 1944.
3. Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale.* New York: Psychological Corp., 1955.

PSYCHOLOGICAL TEST REVIEWS

Edwards, Allen L. *Edwards Personal Preference Schedule* (*PPS*). New York: Psychological Corp., 1953, 1954.

Reviewed by

J. Richard Wittenborn

Rutgers University

The *Personal Preference Schedule,* a thoughtfully constructed instrument, embodies merits expressive of contemporary psychology. Although the author emphasizes the value of his schedule as an instrument for research, it comprises features which may prove to be of great merit in practical assessment.

Based on Murray's system of manifest needs, it provides scores for: *1.* Achievement, *2.* Deference, *3.* Order, *4.* Exhibition, *5.* Autonomy, *6.* Affiliation, *7.* Intraception, *8.* Succorance, *9.* Dominance, *10.* Abasement, *11.* Nurturance, *12.* Change, *13.* Endurance, *14.* Heterosexuality, and *15.* Aggression. The need scores emerge from a procedure whereby the subject indicates his preference in each of a series of 225 pairs of experiences, activities, or situations. These pairs are provided by a matching of each of 15 statements (expressive of one of the needs) with every other one. The selection of the statements to be paired was guided by consideration of the social desirability of the statements. It was hoped that having the members of the pairs equal with respect to social desirability would reduce the probability of the need scores being confounded with their social desirability.

Although the paired-comparison structure of this instrument may control social desirability and reduce the effect of other irrelevant or confounding needs and biases, it may involve possible disadvantages. For example, the subject must express a preference for one of two modes of response both of which may be equally disliked and, in addition, the intensity of the need may remain obscured. It is therefore possible that a given statement of manifest need which is consistently highest in this set of 15 may not represent a truly intense need for the individual. Perhaps, however, it is too much to expect an instrument which shows the relative intensity of needs also to show their absolute intensity.

Both split-half and retest reliabilities are given for the various scores. The corrected split-half correlations range from .60 to .87 with a median of .77, and the retest correlations range from .74 to .88 with a median of .78. Norms, provided for both men and women college students, are expressed in terms of *T* scores as well as percentile ranks. The 1509 cases comprising the standardizing group were drawn from many different colleges broadly representative of geographical areas and types of institutions.

An interesting feature of the PPS is the fact that it yields scores which are relatively independent despite the fact that they are relatively reliable. The highest subtest intercorrelation, .46, is between Nurturance and Affiliation, certainly no more of a relationship than would be anticipated from the similarity of the constructs represented. These two variables, possibly with a third variable, Abasement, seem to bear a small but systematic, negative relationship with such variables as achievement, autonomy, and dominance. These and other possible inconsistencies which may be inferred among the intercorrelations are plausible and do not detract from one's confidence that the variables may be usefully expressive of the need achievement system from which they draw their labels.

The *Personal Preference Schedule* is a modern instrument in the particular respects that it is designed to measure personality constructs which have a conceptual origin independent of the instrument itself; the subtests are not only reliable but relatively independent so that they are free from a confusing and uneconomical overlapping implication; and there is a built-in estimate of the reliability of the individual's performance in the form of a self-consistency score. The user also enjoys substantial protection from the possibility that the manifest need scores could be confounded to an important degree by the need to appear socially desirable. The possibility of personality schedule scores being confounded with a tendency to present a socially desirable facade is often considered to be a fatal defect among such instruments. Edwards not only minimizes this possibility by pairing the statements on the basis of their predetermined social desirability, but he scrupulously examines the degree to which he has succeeded in doing this by the use of other scales and inventories. Incidentally, it is possibly the elimination of the confounding with social desirability which permits the relatively low intercorrelations between the scales.

When one thinks of the validity of an instrument traditionally, one thinks of the degree to which its scores are related to some explicit, practical criterion for some definite sample. Such relationships, however, are but an expression of a possible consequence of validity; they are not the basis for validity. A certain type of empirical validity may be relatively easy to show for most aptitude or achievement tests. Most such tests comprise a set of items or performances which is in effect a sample from a hypothetical population of performances, and the sample of performances comprising the criterion may also be conceived as drawn from the same hypothetical population of performances. This "sampling" approach, which characterizes most of our testing efforts, should require no more than a reliable *test* sample of responses and reliable *criterion* sample of responses in order for empirical validity to be demonstrable. Edwards' instrument should not be confused with this tradition in testing. The *Personal Preference Schedule* is intended to provide a reliable and convenient means for determining the relative intensity of 15 fairly well-known and previously explored motivational systems. It would be very desirable to show relationships between the various manifest need scales provided by the PPS and other measures for these needs, or manifestations of behavior relevant to them. Nevertheless, it is not necessarily incumbent upon the author to provide such relationships in order to claim that he has created a *Personal Preference Schedule* based on the need system.

The Personal Preference scales have been correlated with other scales and inventories and the results do not detract from the hypothesis that the scales do distinguish between individuals in a manner appropriate to the constructs of need which provide the labels. "Validity" in the pragmatic sense cannot yet be claimed for this instrument, however.

Apparently, there are numerous studies under way which employ this instrument and we shall learn about the possible ways in which Edwards' scales for manifest needs are related to a wide variety of behaviors. In a recent communication with the author, it was learned that the published literature now includes 5 references to the *Personal Preference Schedule* and that 32 additional inquiries which are not yet reported involve its use. It would appear, therefore, that the prospects for a psychologically relevant, conceptual validation of this instrument, and of the need constructs on which it is based, are very good. A useful construct of the behaviors validly expressed by a test may eventually have much more practical value than a psychologically indefinable, conglomerate test which is shown to be empirically valid for some one specific criterion of useful behavior.

At the present time this reviewer finds only two aspects of the total procedure which may require some immediate attention. First, the relative homogeneity or heterogeneity of the scales with respect to the intercorrelations among their component items has not been examined; this could be important. Second, the answer sheet, although faultlessly designed for the rapid scoring and collating of results, is a little tedious for the person who is taking the test. From the standpoint of scoring and interpretation, it would be very difficult indeed to conceive of a more convenient and agreeable procedure. The schedule is virtually self-scoring and the profile on the reverse of the answer sheet is self-interpreting.

Edwards, Allen L. *Edwards Personal Preference Schedule* (*PPS*). New York: Psychological Corp., 1953, 1954.

Reviewed by

John W. Gustad
University of Maryland

In developing the *Personal Preference Schedule*, Edwards was seeking ". . . to provide quick and convenient measures of a number of relatively independent *normal* personality variables." Further, the instrument was intended to be useful primarily in counseling and in research. Both objectives are highly desirable. Far too little attention has been paid to the development of measures relevant to the normal range of behavior.

The 15 scales of the PPS ostensibly measure the manifest needs proposed by Murray. There is an additional score, a measure of consistency, which was developed to provide information relative to the subject's test-taking behavior. In addition to its applications to counseling and research, Edwards suggests that the schedule will be useful for classroom demonstrations, in selection problems, and in inverted factor analysis. If its usefulness for these purposes can be demonstrated, the PPS will be a significant addition to the psychologist's armamentarium.

The coincidence of the dates of publication of this instrument and of the APA's *Technical Recommendations for Psychological Tests and Diagnostic Techniques* suggested a preliminary method for assessing its adherence to good practice in development. A check list was developed containing all items in the APA pamphlet which were both relevant to this kind of instrument and which were marked essential. The test manual was then compared with these points, and a decision was reached as to whether the PPS conformed to the essential recommendations. In 54 out of 60 cases it did conform, giving it a "score" of 90 per cent. On the face of it, this would seem to be a favorable sign. Unfortunately, the "score" did not tell the whole story. The percentage of "hits" varied from 100 on recommendations regarding administration to 16 on validity.

The major sections of the manual provide a helpful framework for discussion. The section on development is given over largely to a discussion of the method followed for reducing the effects on responses of the factor of social desirability. Evidence is cited later, in the section on validity, which indicates that this attempt was largely successful. Missing from the section on development is information as to the sources of items and the methods of item selection. The reader is told that each of Murray's 15 needs is paired with each other twice, yielding a total of 225 items. It would be interesting to know how the items were selected and built.

Directions for scoring and administration are fairly simple and straightforward. The PPS is untimed and requires about 40 minutes' working time for most college students for whom it was built. It is not quite self-administering, but chances for error seem small. Scoring is simple and can be done by clerks. Only when the omission of items requires prorating is there any likelihood of difficulty, and this hazard is small. Raw scores may be easily plotted on a profile based on *T* scores; percentile equivalents are also available.

Norms are based on 1509 cases—760 men, 749 women—all students drawn from a number of arts colleges. No information is given regarding the conditions under which the students took the PPS. There is a mean difference of 1.69 years between the sexes in favor of the men but separate norms for the sexes are provided.

Users of the PPS are advised to determine for themselves the limits for high and low scores. Edwards provides tentative cutting points with *T* scores of 70 and above being high, 30 and below being low, the intermediate points being 10-point *T*-score ranges.

As mentioned, the validity check or measure of test-taking attitude is provided by the consistency score, a measure ". . . based on a comparison of the number of identical choices made in two sets of the same 15 items" (*Manual,* p. 6). Edwards shows that the distribution of consistency scores departs from that based on the expanded binomial and concludes that it is effective. He suggests a consistency score of 9 or less be considered grounds for questioning the subject's test-taking attitude. Further evidence on the same problem is based on coefficients of profile stability. Ninety-three per cent of these exceeded chance in the standardization group.

Two kinds of reliability estimates are provided: coefficients of internal consistency (split-half) and of stability (test-retest). The former are reported after correction by the Spearman-Brown formula; the uncorrected coefficients are not reported. The stability coefficients are based on a sample of 89 University of Washington students, a surprisingly small group for such a purpose. The interval between tests was one week. The internal consistency coefficients range from .60 for Deference to .87 for Heterosexuality; the stability coefficients range from .74 for Achievement to .88 for Abasement, certainly as high as those reported for most similar tests. Whether they should be higher is a question imbedded in reliability theory as it applies to personality measures. There is a question about the usefulness of stability coefficients based on a week's interval; it is to be hoped that further work will include longer time intervals and a larger sample of subjects.

Intercorrelations among the 15 scales are presented. Of the 105 entries, 96 are statistically significant. The majority are negative. The highest intercorrelation reported is .46; most are less than .20. It would appear that, by and large, Edwards had succeeded in producing an instrument with relatively independent scales. Since construct validity is involved, a factor analysis would certainly seem to be a requisite next step. It should have been done and reported in the manual.

For all tests, validity is the ultimate question. Here the PPS is most deficient, at least insofar as it is possible to tell from the published manual. A list of research projects in process is included, but, in the absence of information about the results obtained, this list is of limited interest. Unless satisfactory validity can be demonstrated, all other features of a test tend to be meaningless.

The APA *Technical Recommendations* describes four kinds of validity: content validity, predictive validity, concurrent validity, and construct validity. The PPS manual deals only with the last of these. As indicated above, Edwards was seeking to develop measures for the needs described by Murray. It would seem necessary, therefore, that he demonstrate that his scales do in fact measure these needs. No such evidence is presented.

Admittedly, construct validity is difficult to establish but the difficulty in no way reduces the necessity for establishing it. The recommended approaches to construct validation are contained in the APA pamphlet as follows:

"Essentially, in studies of construct validity, we are validating the theory underlying the test. The validation procedure involves two steps. First, the investigator inquires: From this theory, what predictions would we make regarding variations of scores from person to person or occasion to occasion? Second, he gathers data to confirm these predictions" (*Technical Recommendations,* p. 14). It is also suggested that factor analysis is one method for establishing construct validity.

One part of the section on validity contains reports of studies based on self-ratings by subjects. Although correlational situations are implied, results are presented verbally as follows: "The self-rankings of some subjects agreed perfectly with their rankings on the PPS" (p. 13). Unless one paid proper attention to the word "some," he could infer a *rho* of 1.0. What he might conclude after noting the use of the word "some" remains in doubt. Edwards concludes this section of the manual with the following statement: "It is not clear, however, how even perfect agreement between self-rating and inventory scores could be interpreted as bearing upon the nature of the variable being measured by the inventory" (p. 13). In view of the lack of clarity, the omission of this section might have been desirable.

The principal line of validation rests on correlating PPS scores with scores on the Guilford-Martin and the Taylor Manifest Anxiety Scale. The Cooperativeness, Agreeableness, and Objectivity scales of the former were used, based on a total of 106 cases.

With the Guilford-Martin, the PPS scales correlated significantly in 4 cases with Cooperativeness, 11 cases with Agreeableness, and 2 cases with Objectivity. There were only 2 significant correlations with the Taylor. One conclusion which might be drawn from the above is that the PPS measures Agreeableness or lack thereof.

The comments above suggest a dilemma. Either Edwards has based his claims for the validity of the PPS on construct validity or on some other kind. Since the manual contains nothing to support claims of any other kind of validity, we must consider his efforts in regard to construct validity. He has correlated his scales with four others, drawn from the Guilford-Martin and the Taylor. The significant correlations are not interpreted to bear on the construct validity of the PPS. That is to say, no reasons are given to support the notion that the PPS does in fact measure the manifest needs proposed by Murray. It would seem, therefore, that the only conclusion to be drawn is that no usable information is presented regarding the validity of the PPS.

It is the responsibility of the test author and the test publisher to establish the validity of any instrument made available commercially for use in one or several practical situations. In the case of this instrument, both the author and the publisher seem to have fallen short of meeting their responsibilities in this respect. One single, simple step might have been—and still might be—taken to correct this. Across the front of each test and each manual, there should be stamped, in large, red letters (preferably letters which will glow in the stygian darkness of the personality measurement field) the word *EXPERIMENTAL.*

It is experimental. It is an intriguing, promising, in many ways very carefully conducted experiment, but it is still an experiment. Until its validity has been established, it must remain an experiment, and it should not be released for any other purpose.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Psychotherapeutics: Success or Failure

G. R. Pascal and Melvin Zax[1]
University of Tennessee

Whether or not behavioral change can take place concomitant with procedures instituted in the psychological clinic is a matter of belief. Belief is a subjective matter and the nature of the evidence required to achieve it seems to vary from person to person. Thus, aborigines have been found who do not believe that a relationship exists between sexual behavior and pregnancy in the human female. No amount of statistical significance can convince them of this relationship. The facts do not fit into their theoretical framework. In our culture, there are individuals who believe that a person behaves the way he does because "he is possessed of the devil," or because, for some supernatural reason, he is "meant to be that way." Persons with such a frame of reference will not be convinced by the nature of available evidence bearing on behavioral change in clinical subjects. Other more sophisticated persons reject evidence on statistical grounds, implying that no one has "proved" that behavioral change in clinic subjects is a result of clinical procedures. They are, of course, correct. No one has "proved" it in that sense. The best available published evidence bearing on this matter is that recently put forth by Rogers and Dymond (4) in which behavioral change as a criterion variable is largely neglected. Satisfactory evidence, on statistical grounds, does not now exist. Without an established criterion, without the possibility of efficient controls, it is improbable that a rigorous scientific experiment can be conducted in this area at this time.

Evidence, however, does exist. Belief in it will depend on the frame of reference of the individual. Certain biases can be singled out which seem to be related to the acceptance of the evidence available. One is a conviction that the human organism learns. This bias is one to which all of us will give lip service; lip service because, although we believe in the learning of reading, writing, and arithmetic, we do not, many of us, believe that other habits such as attitudes can be learned. Another one seems to be the conviction that human learning can be controlled, that is, that conditions can be set up which will foster or hinder the learning of specific pieces of behavior. A third might be the belief that the human organism has the capacity to react with what it has learned in one situation to another similar situation. With these beliefs, one is in a fair way to accept the evidence bearing on behavioral change in clinical subjects as a result of or concomitant with certain procedures in the clinical laboratory. The available evidence runs somewhat as follows:

A subject is brought, or comes under his own power, to our clinical laboratory. It is determined, on grounds not relevant here, that the purpose of laboratory procedures is to obtain a decrement in the number of hours the subject spends alone. Whereas the subject now spends, on the average, a total of ten out of seventeen waking hours alone, it is desired that he spend, on the average, less than five hours a day alone. Number of hours spent alone becomes our dependent variable. From a study of the subject's past history and whatever we can learn from psychological tests, we formulate hypotheses concerning the intervening variables related to the dependent variable. Since we cannot observe or change hours spent alone outside the laboratory, we need to change behavior in the laboratory that is highly correlated with our dependent variable. Thus, in this particular case, we might set up as intermediate dependent variables such things as number of words uttered by the subject, the amount of time spent by him facing the experimenter, the changes in pitch of his voice (affect), random movements, posture, facial expression, etc. We now set up a controlled stimulus situation in the clinical laboratory, which in interaction with the intervening variables, is calculated to affect the dependent variable.

---

[1] The authors wish to express their gratitude to Drs. K. R. Newton, C. N. Sipprelle, and C. H. Swensen, members of the Staff of the Psychological Service Center, who assisted in the preparation of this paper.

The behavior in question, number of hours spent alone, can vary in three directions. It can go up, down, or remain stationary. Our *prediction* is that it will go down. From the point of view of our prediction, it can only vary in two directions; it can go down or not go down. Disregarding all other matters except those under immediate discussion, the probability of effecting behavioral change in the required direction is one over two for the single case.

We now institute our laboratory procedures and obtain behavioral change in the predicted direction. We get another case with a different, gross, dependent variable, inability to care for her infant in a young mother, and again obtain behavioral change in the predicted direction. With other cases and other dependent variables we, again and again, obtain gross behavioral change in the laboratory that we "guess" is highly correlated with behavior outside the laboratory. Since these are independent events, we can multiply probabilities. What level of significance will satisfy us, .01, .001, or .0000001?

With this orientation, the writers entered the files of the Psychological Service Center of the University of Tennessee. This Service Center is operated by the Department of Psychology for training and research purposes in connection with the Clinical Psychology Training Program. It takes referrals of children and adults from agencies, physicians, the school systems and hospitals, to mention some of the referral sources. The Service Center takes all cases for diagnosis or treatment which are felt to be within the province of a clinical psychologist working in an outpatient setting.

The Service Center has been in operation in its present form since 1951. During that time there have accumulated 1,022 case files. The files were entered with the aid of a table of random numbers. Prior to entering the files, however, certain rules for selection were set up. All cases were used which met the following criteria: (*a*) any case which was terminated by consent of both therapist and patient, (*b*) any case of long-term, "deep" therapy which was in progress or was interrupted, where the patient was seen in at least one hundred interviews (called Type III therapy), (*c*) any case of supportive brief therapy which was in progress or was interrupted, where the patient was seen in at least thirty sessions (called Type II therapy), (*d*) any case where environmental manipulation was the major goal, where the patient was seen in as many as five interviews (called Type I therapy). These criteria were set up on the basis of considerable previous experience. It was felt that cases which had less than the criterion number of interviews for the type of therapy indicated did not constitute an experiment. The criteria concerning what did or did not constitute an experiment in this context are outside the range of this discussion but are analogous, in the reasoning behind them, to the criteria applied in other more systematic experimental settings. In any case, these rules were set up before entering the files. Cases were excluded in which large environmental effects beyond any possibility of control by the therapist interfered with therapy. In this manner ten cases were selected from the files.

For each case, notation was made, from the case history, of behavior which was considered deviant and it was desired to change, as reported either by the patient or someone else who gave the history, the latter very often being the case with child patients. Thereafter, after conclusion or interruption of therapy, a final summary of the case, progress reports, or statements of the patient were used to make a judgment as to whether or not the behaviors previously noted changed in the predicted direction. No follow-up was used, no attempt was made to find out whether or not the patient maintained the behavior change over a period of time. We hold no brief for the objectivity of this procedure; we are cognizant of the possible biases of whoever wrote the case records, of the intangibles underlying the patient's "good" reports, etc. The reliability of judgments as to whether or not behavior changed in the predicted direction, is perfect. This finding is understandable because, disregarding the reliability of the case histories themselves, the questions asked of the case records could be answered by a simple yes or no.

Table 1 shows the results obtained, ten hits and no misses. Skepticism seems appropriate for these data. Therefore we pulled out another ten cases meeting the criteria for selection. Table 2 shows the results, eight hits and two misses. Although expansion of the binomial indicated a satisfactory level of sig-

Table 1

First Group of Ten Meeting the Criteria for Selection

File No.	Age	Sex	No. of interviews	Behavior to be changed	Change	Predicted direction
322	30	F	194	1. Fainting spells 2. Overweight	1. None for 2 years 2. Lost 14 pounds	Yes
355	7	M	50	1. Fearful, withdrawing behavior in school	1. Active participation in school activities, improved grades	Yes
155	27	F	70	1. No social contacts with men 2. No contacts with parents	1. Began dating regularly 2. Began visiting parents	Yes
402	28	F	113	1. Unable to care for children and perform household duties 2. Few social contacts	1. Caring for children and doing housework consistently 2. Began regular church attendance	Yes
302	8	M	34	1. Submissive and passive behavior, e.g., beaten by younger children 2. Frequent accidents	1. Began having fights in school and took responsibility for younger sibs 2. None while in treatment or immediately after	Yes
727	51	F	13	1. Uncontrollable weeping 2. Insomnia	1. Ceased completely 2. Sleeping well	Yes
143	35	M	102	1. Multiple phobias 2. Frequent hostile outbursts 3. No dating	1. These subsided almost completely 2. Ceased completely 3. Frequent dating	Yes
69	27	M	135	1. Drinking excessively to relieve anxiety 2. Somatic symptoms, e.g., stomach cramps 3. Frequent prostitutes	1. Not drunk for 10 months 2. None for one year 3. None for one year	Yes
108	26	M	214	1. Infrequent dating 2. Ashamed of poor hearing and refused to wear a hearing aid 3. Few social contacts	1. More dating 2. Bought hearing aid and used it 3. Began to join clubs, etc.	Yes
54	28	M	41	1. Frequent quarrels 2. No plans for future and dissatisfied with job	1. These greatly diminished 2. Entered graduate school with plans	Yes

Table 2

Second Group of Ten Meeting the Criteria for Selection

File No.	Age	Sex	No. of interviews	Behavior to be changed	Change	Predicted direction
380	47	M	10	1. Acute alcoholic episode where he was continually drunk	1. Sobered up	Yes
				2. Unable to work	2. Working steadily for 2 years	
369	42	M	18	1. Using narcotics	1. None for 1½ years	Yes
				2. Crying spells	2. None for 2 years	
				3. Hypochondriacal complaints	3. None for 2 years	
884	5	M	60	1. Unintelligible speech	1. Now easily understood	Yes
				2. Poor coordination (muscular)	2. Better coordination	
				3. Not completely toilet trained	3. Toilet training completed	
				4. Played with electrical outlets and received severe shocks	4. No shocks for 10 months	
384	19	M	162	1. Several attempts at suicide	1. None for 2½ years	Yes
				2. Crying, tremulousness	2. None for 2½ years	
				3. Poor grades at college	3. Better grades	
184	32	M	315	1. Took too much insulin for diabetes against doctor's orders	1. This corrected	Yes
				2. Hypoglycemic attacks	2. None for 3 months	
				3. Frequent awakening at night	3. Ceased	
				4. Very frugal	4. Now buys furniture, etc. and takes trips	
				5. Very passive in dealing with subordinates	5. Has been able to fire subordinates when necessary	
974	32	M	25	1. Had failed doctoral language exam twice	1. Passed language exam and prelims later	Yes
				2. Verbally threatened violence toward a professor	2. Stopped	
865	21	F	48	1. Few close friends	1. No change	No
430	29	M	40	1. Overt homosexual acts	1. No homosexual acts for 8 months	Yes
				2. Frequented prostitutes	2. None for one year	
				3. Constant extreme muscular tension	3. Moderate tension occasionally	
292	38	M	252	1. Inhaled nitrous oxide and induced hallucinations	1. Gradual decrease until total stop 6 months ago	Yes
				2. Sat up nights listening to Bible on tape recorder	2. Ceased	
171	22	M	31	1. Excessive drinking	1. This reduced	No
				2. High blood pressure, somatic weakness, burning sensation in stomach	2. Became more severe and hospitalization became necessary	
				3. Hyperactive	3. No change	

nificance even for this finding (.05) we were nevertheless worried by the trend indicated in the two misses obtained. We selected another ten folders from the files. Table 3 shows the results, ten hits and no misses. The finding satisfied us. We felt we could safely bet on the side of the "hits."

Tables 4 and 5 show all of the cases pulled from the files, 350. It will be noted that the bulk of these were diagnostic only. The category "Inquiry Only" exists because any first interview conducted at this Center is usually written into the files despite the fact that the inquiring person may never return to the Center. The second largest category listed in the table is that labeled "Too Few Sessions."

Table 3

Third Group of Ten Meeting the Criteria for Selection

File No.	Age	Sex	No. of interviews	Behavior to be changed	Change	Predicted direction
405	8	M	47	1. Failing grades at school 2. Frequent disciplinary problem	1. Began receiving passing grades 2. Marked reduction in misbehavior	Yes
895	24	F	5	1. Indecision over impending engagement 2. Dreams and nightmares 3. Had broken engagements in past and avoided becoming serious with boys	1. Became engaged 2. These ceased 3. Got married	Yes
14	29	F	197	1. Suicide attempt 2. Constant sitting on husband's lap 3. Weeping behavior 4. Few social contacts	1. No further attempts 2. Ceased 3. Ceased 4. Began attending social affairs often	Yes
385	4	M	163	1. No speech 2. No response to others 3. Assumed positions, posturing 4. Habitually constipated	1. No change 2. Attempts to communicate by pointing 3. This done infrequently 4. No problem for 6 months	Yes
616	25	F	44	1. Few social contacts 2. No pleasure in sex relationships 3. Fear of driving a car	1. Joined clubs, etc. 2. Began to achieve orgasm 3. Learned to drive	Yes
67	32	M	93	1. Wife beat him 2. Spells of extreme physical weakness 3. Crying spells	1. This ceased and at times he struck her 2. None for about 3 years 3. None for about 3 years	Yes
110	22	M	25	1. Attacks of nausea and depression 2. Left school and was indecisive about future	1. Ceased for about 3½ years 2. Returned to medical school	Yes
226	20	M	65	1. Failing courses at school 2. Not eating and sleeping well	1. Improved grades and graduated 2. Eating and sleeping well	Yes
170	32	M	155	1. Worked over-long hours 2. Unable to communicate to wife or subordinates at work	1. Ceased 2. Easier verbalizations with people	Yes
325	22	F	77	1. Excessive sleeping 2. Exhaustion and nausea often 3. Hypnogogic experiences 4. No dating	1. This ceased 2. This ceased 3. These remained 4. Got married	Yes

This category is probably the one that would be most suspected in estimating the effects of psychotherapy in our presentation. It will be noted that there are more cases in this category than the thirty that were accepted for the study. Whether or not these can be considered failures in psychotherapy is a matter for discussion beyond the realm of this paper. We feel that the matter is irrelevant to our main purpose. The category "Environment Effect" is broken down and self-explanatory. Of the thirty cases retained for the analysis, the table shows that there were five children and twenty-five adults. Most of the adults were in what we call Type II therapy which consists of attempts to modify variables within and without the person. Type I therapy is simply environmental manipulation. In Type III therapy an attempt is made to concentrate on variables within the individual to effect behavioral change with relatively little attention being paid to the environment.

Obviously, the tables do not "prove" anything. It is hoped, however, that they might contribute to belief and, more importantly, illustrate the possibility of a systematic approach to the problem of success or failure in psychotherapy. It should be pointed out that we have not been concerned with whether the patient was "improved" or "unimproved" as has been done in previous studies (1, 2). It is impossible to see how any judgment can be made, let alone refined statistical procedures applied with unreliable data such as "improved" or "unimproved." Nor have we been concerned with whether or not the mechanisms of ego-defense became more suitable (3) or if the patient was more comfortable or optimistic (4) or a host of other similar measures proposed to estimate the effects of psychotherapy.

Table 4

Analysis of Cases Rejected

Reason for rejection		Number
Diagnostic only		231
Inquiring only		34
Environmental effect		12
Couldn't pay	5	
Hospitalized	3	
Left town	3	
Withdrawn by physician	1	
Too few interviews to meet criterion		43
		—
Total cases rejected		320

Table 5

Analysis of Cases Retained

Type of therapy	Mean No. of Hours Seen	Number	Behavior changed in predicted direction	
			Yes	No
Type I	10	1	1	0
Type II	60	16	14	2
Type III	184	8	8	0
Play	71	5	5	0
		—		
Total cases retained		30		

By all means let us make the patient's ego-defenses more suitable. Let us change his attitudes about himself and others. But these have no significance if they are not reflected in the behavior of the patient as it is displayed in the environment in which he has to live. Making ego-defenses more suitable (or what have you) may be an hypothesis concerning a *way* to effect behavioral change but it cannot be used as a criterion. Neither can unvalidated tests and questionnaires (no matter how much they are Q-sorted) be used as criteria. If the ultimate purpose of psychotherapy is to effect behavioral change (and it has to be) then *behavior* must be the criterion. This proposition not only makes sense but it gets us closer to the science of psychology.

The writers are aware that something other than mere behavioral change is considered in the problem of success or failure in psychotherapy. There is, for instance, the consideration of time during which behavioral change is maintained after psychotherapy. The answer to this question pushes us back to our systematic formulation. If behavior is changed by simple environmental manipulation (Type I) then the maintenance of changed behavior will depend on the stability of the changed environment. If, on the other hand, the process of changed behavior was accomplished exclusively by manipulation of variables within the individual (Type III), then the changed behavior should be maintained with relative freedom from the environment. No thoroughgoing study with results of psychotherapy can neglect a consideration of procedure used to effect behavioral change if the time dimension is to be added to the notion of success or failure. Another consideration is that, although success or failure may be concerned with the achievement, in the environment, of a desired and predicted change in behavior, the question as to what behaviors, and how much they have to be changed, is unknown at this writing. This consideration leads us into the morass known as "psychological health." We shall by-pass this matter and end our polemic by asserting that we believe a person's behavior can be changed in a predicted direction by the systematic use of well-established principles known to the science of psychology.

Received May 28, 1956.
Early Publication.

References

1. Eysenck, H. J. The effects of psychotherapy: an evaluation. *J. consult. Psychol.*, 1952, **16**, 319–324.
2. Landis, C. Statistical evaluation of psychotherapeutic methods. In L. E. Hinsie, *Concepts and problems of psychotherapy*. New York: Columbia Univer. Press, 1937. Pp. 155–169.
3. Miller, J. G. Criteria and measurement of change during psychiatric treatment. *Bull. Menninger Clin.*, 1954, **18**, 130–137.
4. Rogers, C. R., & Dymond, Rosalind F. *Psychotherapy and personality change*. Chicago: Univer. of Chicago Press, 1954.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Relationship of Success and Length in Psychotherapy

James W. Taylor
Veterans Administration Hospital, Denver, Colorado

Cartwright (1) has suggested the existence of a "failure zone" between the 13th and 21st interview in client-centered therapy. The present report is made in order to indicate a similar result occurring in a psychoanalytically oriented clinic. In the process of other research, the total number of cases closed in the clinic during a fiscal year (309 cases of varying diagnosis and with different therapists) were evaluated for number of treatment interviews and for results of treatment. Results of treatment are evaluated as, 0 unimproved, 1 mildly, 2 moderately, and 3 greatly improved by the therapist and consultant. While this criterion for improvement leaves much to be desired, such judgments are common in most clinics operating under practical restrictions. Figure 1 combines the data presented by Cartwright in three graphs with the corresponding data of this clinic. Representation of the two scales of improvement on the same figure was not meant to imply their equality; rather, the purpose is to illustrate the similarities of the curves.

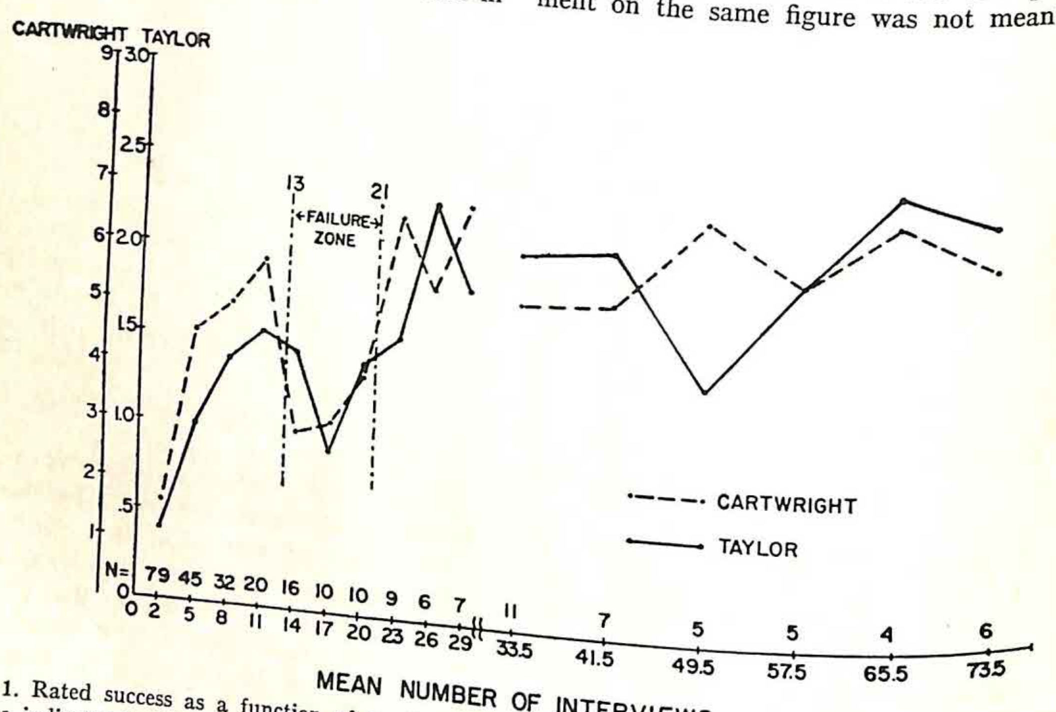


Fig. 1. Rated success as a function of number of interviews. The break in the horizontal axis at 30 interviews indicates that both the class intervals of 29 and 33.5 include the 30th interview (4 cases) which was necessary in combining the graphs. *N* indicates the number of patients in each class interval.

Received January 13, 1956.

Reference

1. Cartwright, D. S. Success in psychotherapy as a function of certain actuarial variables. *J. consult. Psychol.*, 1955, **19**, 357–363.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

The Representation of Orthopedic Disability in Children's Figure Drawings

A. B. Silverstein and H. A. Robinson
Psychiatric Institute, University of Maryland School of Medicine

Twenty years have passed since Schilder first expressed his belief that "the way in which children draw human figures really reflects their knowledge and sensory experience of the body image" (14, p. 106). Since that time many investigators have developed this observation both at the level of theory and in clinical practice. Body image has been variously described by different writers, and the lack of a universally accepted definition has undoubtedly served as a deterrent to systematic research in this problem area. Nonetheless there has remained a widely held belief that drawn human figures do reveal something of the subject's conceptual picture of his own body. Machover, the most active proponent of dynamic interpretations of figure drawings, has stressed that "the process of drawing the human figure is for the subject . . . a problem not only in graphic skill, but one of projecting himself in all of the body meanings and attitudes that have come to be represented in his body image" (10, p. 35). Today, figure drawings are a commonly accepted tool for investigating personality dynamics, self concept, and body image; a familiar part of the clinical battery. This technique, however, shares with many other projective instruments the need for convincing and unequivocal demonstrations of validity.

The present authors are engaged in an investigation of psychological factors in the treatment of and recovery from poliomyelitis,[1] and have had the opportunity to study figure drawings of children in various stages of that illness. In exploratory phases of this work we were frequently impressed with what appeared to be a rather direct reflection of the child's disability in his drawings, and, given impetus by these observations, decided to initiate a more systematic study of the technique. If this procedure could be demonstrated to yield a true picture of body image, it would prove extremely useful, since data from other sources than figure drawings have shown that the discrepancy between body image and the actual physical body is related to adjustment processes in physical disability (2). It should also be noted that figure drawings as such have received scant attention as possible measures of attitudes toward disability. A nonverbal measure of disability-associated attitudes and modes of adjustment would be of great value to workers who deal with the disabled.

The earliest study of the figure drawings of handicapped subjects is reported by Machover (10). Her brief account of this investigation of 20 orthopedic cases provides no methodological details. It would appear that drawings were simply scrutinized for the possible projection of disturbances of body image, and the preliminary results are said to have been striking. Other investigators (1, 6, 12, 16) have taken a similar approach to drawings produced by various types of disabled subjects, and their results, while not as impressive as Machover's, seem to be encouraging. Taken together, these studies suggest that, while relatively few subjects portray their disabilities directly, indirect evidence in the form of distortion and displacement occurs frequently.

---

[1] Aided by a grant from the National Foundation for Infantile Paralysis.

Bureau Ednl. Psy. Research
DAVID HARE TRAINING COLLEGE
Dated ....

1. Study of Subjective Classifications

Method. The writers have used a sorting approach in studying a series of drawings by disabled children. Same-sex, opposite-sex, and self-figure drawings were obtained from a group of 22 chronic poliomyelitis cases with paralytic residuals of the lower extremities. Each of the three sets of drawings was evaluated independently by the authors. With no preliminary formulation of a classification scheme, and with negligible disagreement between the judges, we found that these drawings could readily be sorted into six categories. The distinguishing features of these groupings were: (*a*) omission of a portion of the legs; (*b*) external support for the drawn figure, e.g., by an orthopedic appliance; (*c*) marked size discrepancy of the legs; (*d*) improper attachment of the legs to the trunk; and (*e*) unusual treatment of the upper extremities, with the legs judged to be within normal limits. The final category (*f*) included drawings considered to embody neither direct nor indirect evidence of disability.

Results. Four of the 22 self-figures are presented in Fig. 1. Though the drawings differ a great deal, each appears to show clearly and dramatically some representation of impairment. In one way or another these children seem to have projected their physical condition upon their drawn figures. Nor are

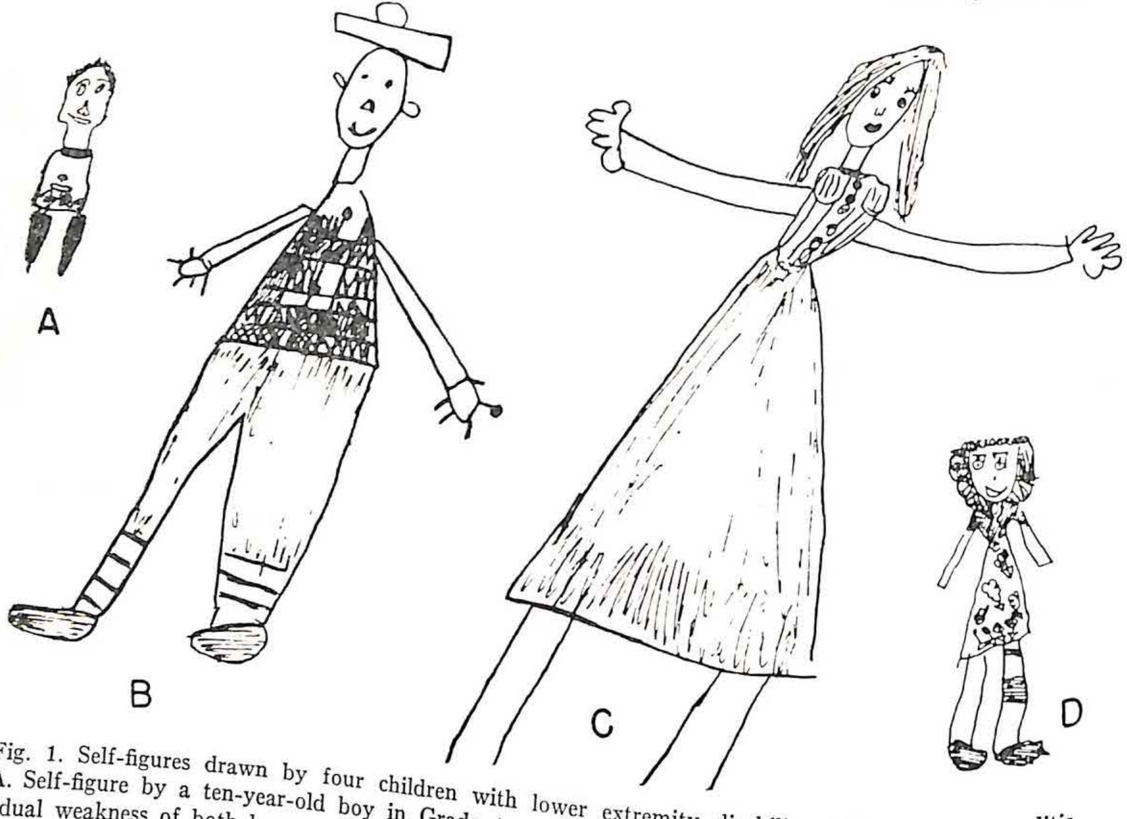


Fig. 1. Self-figures drawn by four children with lower extremity disability following poliomyelitis.

A. Self-figure by a ten-year-old boy in Grade 4. He wears bilateral full-leg braces and has a marked residual weakness of both legs as a result of a spinal poliomyelitis at age two.

B. Drawing by a boy, a nine-year-old second-grader. Disease onset was at age five. Residual is now restricted to the right leg: he wears a short leg brace.

C. Drawing by a nine-year-old girl in the second grade. Since the poliomyelitis onset before her first birthday, she has been readmitted to an orthopedic hospital on five occasions for observation and corrective surgery. Severe paralytic residuals of both legs. She wears a long leg brace with corrective shoes. Note that the legs of her figure are "paper-chopped."

D. Figure by a ten-year-old girl in Grade 3. The disease onset occurred at age 5. Her right leg is essentially flail below the knee, and she wears a long leg brace with pelvic band. This child has had three hospital readmissions for corrective surgery. Note that she has faithfully pictured her brace, though in mirror-fashion.

Table 1

Classification of Figure Drawings by 22 Children With Lower Extremity Disability Following Poliomyelitis

Group	Same-sex	Opposite-sex	Self-figure
Portion of lower extremities omitted	5	3	3
External support provided	0	0	4
Size discrepancy in lower extremities	6	7	5
Improper attachment of legs	3	4	2
Unusual treatment of upper extremities	4	3	4
No evidence of disability	4	5	4
N	22	22	22

these atypical examples; many of the drawings, by direct representation, displacement, distortion, or other unusual treatment, are equally suggestive of body-image disturbance.

Table 1 presents the distribution of figure drawings among the six categories which have been described. In each of the three sets of drawings, more than three-quarters of the subjects were judged to have portrayed a physical disability, either directly or indirectly. By inspection the three sets of drawings show similar distributions among the categories, with the one exception of the "external support" classification. In this instance three self-figures were depicted with orthopedic appliances, and one seated at a desk. This might suggest that the most direct representation of disability may be found in self-figures.

An additional observation should be made with respect to a pattern of consistency found in these drawings. If an individual omits a part of the legs of one of his figures, for example, he tends to treat the other figures similarly. Of the 22 subjects, 16 have either two or three of their drawings in the same category, whereas by chance this should occur only 9 or 10 times. This difference from chance expectancy is significant at the .01 level.

Discussion. While these findings, like those of the investigations cited above, appear very encouraging, it is clear that this subjective approach to the study of figure drawings is seriously deficient from the point of view of scientific rigor. For one thing, the judges who evaluate the drawings are not only aware that they are dealing with productions of handicapped individuals, but also have some general knowledge of the nature and site of the disability. This may introduce bias in at least two ways: the observer may tend to overweigh minor deviations as evidence of disability, or he may de-emphasize certain characteristics not obviously related to the known impairment. Another difficulty in this method of dealing with figure drawings stems from the enormous variety of the productions. Even when working with small samples, one must arbitrarily limit the number of classifications to avoid the unmanageable proliferation of categories. And this inevitably poses the need for an equally arbitrary system of priorities, since many drawings will embody the characteristics specified in more than one category. Drawing A in Fig. 1 illustrates this problem: the figure lacks feet, and the legs are attached to the body in an unusual way. An even stronger objection to this approach is that the frequency of occurrence of the various drawing features in a comparable physically normal sample is unknown. To the extent that the drawings of clinically normal individuals are found to exhibit the same characteristics as those of disabled persons, the utility of the technique is diminished.

Taken together, criticisms of this sort seriously undermine any conclusions which may be drawn from studies employing subjective classification and appraisal of human figure drawings. To be sure, normal controls and objective methods have not been entirely neglected in studying figure drawings of the disabled. Several investigators (3, 12, 15) have compared the drawings of handicapped and normal individuals, using checklists of signs or physical measurements of the drawings. Without exception these investigations yielded negative results, in the sense that drawings of disabled and normal subjects were not reliably distinguished. In each case, however, it would appear that flaws in sampling or methodology may have had the effect of obscuring differences in the drawings. In this connection the following points deserve mention: Failure to control for correspondence

Table 2

Comparison of Disabled and Control Subjects with Respect to Age, Intelligence, and Grade

Group	Age		IQ		Grade	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Disabled ($N = 22$)	11.2	1.19	94.6	9.27	4.7	1.48
Control ($N = 44$)	11.1	1.02	97.6	7.26	5.2	0.94
t	0.48		1.40		1.52	
F		1.39		1.67		2.54
p	$>.10$	$>.10$	$>.10$	$>.10$	$>.10$	$<.05$

between the sex of the subject and that of his drawn figure denies, in effect, the possibility of differential identification with, and projection on, same-sex and opposite-sex figures. Inclusion of items in scoring schemes which are not directly related to the known impairment, or which cannot in reason be justified as possible displacements or symbolizations of the disability, may needlessly elaborate the measuring instrument without increasing its sensitivity. An example of a seemingly irrelevant item, reported in the literature, is one dealing with the Adam's apple in a study of the drawings of disabled children. Use of extremely heterogeneous groups—a child with a congenital deformity of an arm in the same sample with a child having an osteomyelitis of the leg, for example—seemingly disregards the psychological significance of the nature and site of the disability.

It should fairly be stated that the heterogeneity of disabled individuals poses a difficult and fundamental problem for any research effort in this area. Besides the sources of variation found in physically normal individuals, such considerations as site, degree, and duration of disability make it virtually impossible to secure properly homogeneous samples of adequate size. Thus the research worker must knowingly and selectively sacrifice such variables as seem least pertinent to his goal, and at the same time remain sensitive to the need for homogeneity. The 22 subjects of the present study can be considered homogeneous to this extent: They share the same disease entity and the same general site of disability. They are enrolled in a public school equipped to deal with orthopedically disabled children who by reason of their handicap cannot attend their regular classes. This is an environmental control of a sort, and also ensures that each subject has a marked degree of disability. Age of onset varies greatly in this sample, but each child was at least a year and a half past the acute stage of the disease when he was seen.

2. Controlled Study of Signs

Method. Having considered these subjects in the light of factors related to their disability, it is now necessary to report some other characteristics of the group, and to describe a sample of normal subjects who were introduced as a control. The disabled children ranged in age from 9 to 12 and were equally divided by sex. Over 100 physically normal children from the same metropolitan school system were tested, and 44 of these subjects were selected to form a group comparable to the disabled sample with respect to age, grade, and reported intelligence quotient (Kuhlmann-Anderson). These 44 nondisabled children, equally divided by sex, also ranged in age from 9 to 12. The control group was established at twice the size of the experimental sample to increase the reliability of comparisons, and because physically normal subjects were easily available. The group of children with poliomyelitis residuals, on the other hand, represents nearly all available cases. Table 2 presents a comparison of the two groups.

The equation of the two groups is seen to be entirely acceptable for age and intelligence and mean grade placement; no significant differences are obtained. The significantly greater variance in grade placement observed in the disabled sample may well reflect differing periods of absence from school during treatment for poliomyelitis. With respect to the measures reported, it is submitted that the two groups are essentially comparable.

Each sample was tested in a classroom situation during the school day. The conventional figure-drawing procedure was followed: the subjects were instructed to draw a person, and a person of the opposite sex from the first. They were then asked to draw pictures of themselves. No time limits were set, and no further directions were supplied.

To analyze the drawings obtained, a list of 55 scoring items or signs was developed. Thirty-seven of the signs deal specifically with types of treatment which may be given to lower extremities in figure drawings; for example, shape of foot, length of leg, and omissions. Six signs call for comparisons between the arms and legs of the figures, as in noting differences in the use of shading. The remaining twelve signs involve no consideration of the legs of the figure, but pertain to other features which might be expected to reflect body-image changes resulting from disability, such as the size of the figure and its placement on the paper. Most of these scoring signs were taken directly or adapted from the reports of Buck (5), Goodenough (8), and Machover (10), and the remainder devised by the present writers.

For each sign, specifications were drawn up to increase objectivity of scoring, and a preliminary test of the procedure was satisfactorily completed. The drawings of the experimental and control groups, combined in a random order, were then scored independently by the authors. Each sign was scored plus, minus, or unscorable. As an example, a figure with misshapen feet would receive a plus score for a particular item; a figure with feet judged to be normal would be scored minus; and a figure from which the feet had been omitted would be considered unscorable for this item. Of the 10,890 individual signs (55 signs for each of three drawings by 66 subjects), 88 per cent were scored in agreement by the two judges. Scoring disagreements were later resolved in conference. For individual signs, the percentage of initial agreements ranged from 54 to 98 per cent, but for only three items was it less than 80 per cent. An agreement of 33 per cent would be expected by chance. On the whole, the obtained scoring reliability is considered to be acceptable.

Results. In statistical analysis, the three sets of drawings (same-sex, opposite-sex, and self-figures) were treated separately. For each set, chi square corrected for continuity was employed to compare the drawings of the two subject groups on the 55 signs. Fisher's method (7, p. 96) was used to determine the exact probability associated with chi squares which approached significance. For each comparison, the null hypothesis was tested against the alternative that the disabled group contained a greater proportion of subjects scoring plus on a given sign. Table 3 presents a summary of the statistically significant differences found among the 165 tested.

In the 55 comparisons made for each of the three sets of drawings, two or three dif-

Table 3

p Values Exceeding the Five Per Cent Level in Comparison of Figure Drawings of Disabled and Control Subjects

(Total number of comparisons for each figure is fifty-five)

Item	Same-sex	Opposite-sex	Self-figure
3. Figure placed at side edge of paper			.039
16. Horizontal midline (waist area) emphasized	.017		
20. Feet differ in size or shape (front view)		.035	
22. Leg drawn lighter, thinner, or fainter than corresponding arm		.049	
23. Figure placed above horizontal midline of paper	.014		
26. Length of leg greater than twice length of trunk		.037	
28. One leg thinner than the other			.029
51. Width of leg increases toward foot		.039	
55. Leg attached to but not continuous with body		.0002	.009

ferences significant at the .05 level or better might be expected on the basis of chance alone. It would be necessary to obtain at least six such differences to rule out the operation of chance factors (at the .05 level) (13). The number of differences actually found to be significant at the .05 level or better, for the same-sex, opposite-sex, and self-figures, was two, five, and three respectively.[2] It is accordingly clear that the drawings of the disabled and normal subjects have not been reliably distinguished by the scoring system employed.

Discussion. The results of this objective and analytic approach are plainly less encouraging than those obtained when the drawings, as described earlier, were sorted into a limited number of categories. It happens that two drawing features are given similar consideration in these widely different methods of appraisal, and it is of some passing interest to scrutinize them. In the initial sorting procedure, malattachment of legs to body was thought to be a rather direct representation of impairment. Among the 55 signs there appears a virtually identical item (No. 55). The sign analysis does in fact show that—for two of the three figures—the disabled group contains a significantly greater proportion of subjects scoring plus on this item. In the other instance, the omission of a part of the leg placed a drawing in one of the original categories, and again this item is duplicated in the list of signs. Here the sign approach does not appear to bear out the original impression, namely, that omitting a lower extremity part is an evidence of disability. In their drawings the handicapped children do not more frequently omit parts of legs than do the controls. On the contrary, a significantly greater proportion of the normal subjects makes such omissions. (The difference, for which p is less than .05, lies in the unpredicted direction, and is therefore not reported in Table 3.) This, clearly, is an example of a finding of a subjective and impressionistic approach which does not stand up under more careful scrutiny.

From the methodological point of view, any subjective classification method is clearly inferior to the objective and analytic sign approach. It should, however, be noted that this latter method is itself not beyond criticism. As mentioned earlier, the bulk of signs used in this study dealt specifically with types of special treatment given to the legs of the drawn figures, with the rationale that the primary disease residual was in the lower extremities of the subjects. Yet it may be that restricting the choice of signs in this way has unduly de-emphasized other distinctive treatments, as when the body-image representation of the physical disability generalizes, or is displaced to other portions of the body than the legs. At a somewhat different level, the criticism has been offered that "the check-list [of signs] violates that holistic, dynamically oriented approach which competent clinicians use in making drawing analyses" (4, p. 179). Criticisms of this kind seem entirely reasonable, and must be taken seriously. It is possible, with the materials of the present study, to utilize a more global, "clinical" approach and at the same time retain a measure of scientific rigor.

Clinical judgments as to whether figure drawings were executed by disabled or normal subjects have been reported in the literature (3, 15). The findings of these studies have been negative, i.e., the drawings of normal and disabled groups were not reliably distinguished. However, some of the same flaws (heterogeneous samples and the failure to control for sex) found in studies involving the sign approach may have also served to obscure differences between groups in investigations using clinical judgments. The relative homogeneity of subjects in the present study, and control of the sex factor, here provide a better opportunity for such differences to appear.

3. Controlled Study of Global Judgments

Method. Each of three judges was given the three randomized sets of same-sex, opposite-sex, and self-figure drawings, with instructions to select—by whatever cues the drawings might provide—the 22 figures in each set which had been produced by the physically disabled children. The three judges were postdoctoral clinical psychologists experienced in the psychodiagnostic use of figure drawings. Before evaluating the drawings, they were given information about the general characteristics of the two subject groups (age, sex, and IQ), and about the special characteristics of the chronic poliomyelitis children (nature, site, and duration of disability). For an individual figure, the judges knew only the sex of the child who had made the drawing; their task, in effect, was to choose—from 66 drawings—11 drawings by boys and 11 by girls which they considered to be productions of disabled subjects.

---

[2] These figures refer only to differences significant in the predicted direction. Three differences which were statistically significant in the unpredicted direction—where the control group exceeded the disabled sample in proportion of plus scores—are not reported in Table 3.

Results and discussion. The choices of each of the three judges were first compared with those of the others, for the three sets of drawings. On the basis of chance, 56 per cent of the judgments of any two clinicians would be expected to agree. (Two judges, who agree by chance on one-third of their choices of disabled children's drawings, will also have implicit agreement on two-thirds of their evaluations of figures by normal subjects.) The actual percentage of agreements between judges ranged from 52 to 79 per cent for the nine possible combinations (three sets of drawings seen by three judges). Four of the comparisons yielded better than chance agreement (.05 level), suggesting that to some extent the judges were using similar cues.

In considering the ability of the judges to distinguish between the drawings of disabled and normal children, 56 per cent of the selections might again be expected to be correct by chance. Correct judgments actually ranged from 52 to 64 per cent for the nine sets of evaluations. No one set of drawings was more accurately appraised than any other, and no one judge was superior to the others in his ability to distinguish among the drawings. Nor were any judge's ratings significantly better than chance.

These experimental findings, like those obtained with the sign approach, do little to enhance the value of the figure-drawing technique for the study of body image in physical disability. The promise of the original impressionistic sorting method has been fulfilled neither by an objective, analytic procedure, nor by a controlled holistic approach.

Discussion and Implications

A number of hypotheses may be suggested to account for the negative results which have been reported, though none of these hypotheses can be tested with the present data. Harris and Goodenough, on the basis of a large-scale investigation of children's drawings, have concluded that it is somewhat doubtful whether children actually do project the body image upon drawn figures. They remark that "so much of the drawing technique in children under 12 or 13 years of age relates to concept formation and understanding, that probably only a small part of drawing behavior relates to the child's interest, temperamental qualities, and other personality features."[3] Since few of the subjects in the present study were older than twelve, it may be that only a small part of the drawing behavior sampled is related to body image.

Levy's exposition of the technique contains some pertinent observations. Noting that "some clinicians interpret each drawing as a projection of the body image" (9, p. 260), he acknowledges that this sort of projection does frequently occur. He nonetheless goes on to point out that a drawing may also represent many other things, e.g., "a projection of attitudes toward someone else in the environment, a projection of ideal self-image, or a result of external circumstances." Levy remarks, moreover, that "it is usually a combination of all of these." Little wonder that the body image is so difficult to tease out of so complex and intricate a production!

A third possible explanation for the negative results stems from an observation by Machover, who writes that "a leg or an arm may be omitted or distorted for reasons of crippling anxieties, just as surely as for the physical crippling of those limbs" (11, p. 263). This is a powerful and compelling argument: if psychological and physical factors are reflected equally and indistinguishably in drawings, any attempt to differentiate the drawings of disabled and normal subjects seems doomed to failure, however rigorously controls are established. Having made this attempt with particular and systematic attention to problems of control, sampling, and methodology, and without securing positive results, we are now inclined toward this pessimistic point of view. The present study and other similar investigations have rested on the assumption that the physical body, the body image, and the drawn figure are in isomorphic relation. Given careful exploration, this one-to-one relationship does not seem to exist.

---

[3] Harris, D. B., personal communication, January, 1955.

We are not prepared, on the other hand, to concede that the figure-drawing technique is without potential value in the study of physically disabled persons. A more individualized approach, focusing on the adjustment patterns of each subject, may bring greater understanding of the widely varying modes of expression displayed by poliomyelitis children in their drawings. For example, to consider those children whose drawings include a clear depiction of their orthopedic appliance—are they, as Machover suggests, "basically body-centered individuals whose whole lives have been blighted by their handicap?" (11, p. 262). Or are they children who have so completely accepted their handicap and incorporated it in their body image, that they feel free to represent it in their drawings? In what ways do such children differ from those who completely omit the legs in their drawings, and from those whose drawings show no evidence whatever of physical disability? In future research designed to answer such questions, we propose systematically to explore these individual differences in terms of psychological processes relevant to physical disability, e.g., defense mechanisms. We are, finally, in complete agreement with Machover's observation that "organs are set in personalities, and have no disconnected life of their own, and . . . the total reaction to handicap . . . is a more vital measure than the specific concrete projection" (11, p. 264).

Summary

Sets of human figure drawings—same-sex, pposite-sex, and self-figures—were obtained rom *22* children in the chronic stage of poliomyelitis. These drawings were studied by three different methods to determine whether they reflected the physical disability of the subjects.

1. By inspection, more than three-quarters of the subjects appeared to represent their disability, either directly or indirectly, in their drawings.
2. Comparison of drawings of the disabled children with those of normal controls, equated for age, sex, and IQ, showed a nonsignificant number of significant differences between the two groups on a series of 55 scoring signs.
3. Judges experienced in the psychodiagnostic use of figure drawings were unable to differentiate the drawings of disabled and normal subjects at a level better than chance.

Various hypotheses were suggested to account for the negative results, and a weakness in the basic assumption underlying the present investigation and similar studies was noted and discussed. Plans for future research, involving a more individualized and personality-centered approach to the human figure-drawing technique, were outlined.

Received January 6, 1956.

References

1. Abel, Theodora M. Figure drawings and facial disfigurement. *Amer. J. Orthopsychiat.*, 1953, 23, 253–261.
2. Bender, Lauretta. Psychoses associated with somatic diseases that distort the body structure. *Arch. Neurol. Psychiat.*, 1934, 32, 1000–1024.
3. Blanchard, Helen M. A preliminary study of the drawing of the human figure by brain injured children. Unpublished master's thesis, Univer. of Denver, 1952.
4. Brown, F. House-Tree-Person and human figure drawings. In D. Brower & L. E. Abt (Eds.), *Progress in clinical psychology*. Vol. I, Section 1. New York: Grune & Stratton, 1952. Pp. 173–184.
5. Buck, J. N. The H-T-P technique. *J. clin. Psychol. Monogr. Suppl.*, 1948, No. 5.
6. Fielding, B. B. Attitudes and aspects of adjustment of the orthopedically handicapped woman. Unpublished doctor's dissertation, Teachers Coll., Columbia Univer., 1950.
7. Fisher, R. A. *Statistical methods for research workers.* (11th Ed.) New York: Hafner, 1950.
8. Goodenough, Florence L. *Measurement of intelligence by drawings.* Yonkers, N. Y.: World Book Co., 1926.

9. Levy, S. Figure drawing as a projective test. In L. E. Abt & L. Bellak (Eds.), *Projective psychology*. New York: Knopf, 1950. Pp. 257–297.
10. Machover, Karen. *Personality projection in the drawing of the human figure*. Springfield, Ill.: Charles C Thomas, 1949.
11. Machover, Karen. Discussion of Theodora M. Abel, "Figure drawings and facial disfigurement." *Amer. J. Orthopsychiat.*, 1953, **23**, 262–264.
12. Prater, G. A. A comparison of the head and body size in the drawing of the human figure by hemiplegic and non-hemiplegic persons. Unpublished master's thesis, Univer. of Kentucky, 1950.
13. Sakoda, J. M., Cohen, B. H., & Beall, G. Test of significance for a series of statistical tests. *Psychol. Bull.*, 1954, **51**, 172–175.
14. Schilder, P. *The image and appearance of the human body*. New York: International Universities Press, 1950.
15. Simms, Nancy. An analysis of the human figure drawings of orthopedic and non-orthopedic children. Unpublished master's thesis, Univer. of Nebraska, 1951.
16. Wille, W. S. Figure drawings in amputees. *Psychiat. Quart. Suppl.*, 1954, **28**, 192–197.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

A Comparison of Productivity and Fantasy on the Rorschach and Projective Movement Sequences[1, 2]

William H. Lundin and Boris Breiger
Chicago State Hospital

Clinicians frequently rely upon inference when analyzing sparse or unelaborated projective records. Clinical utility increases when such records more clearly show psychodynamic content. Experience with the Projective Movement Sequences Test (PMS), a brief motion picture comprising eight semistructured and ambiguous sequences, suggests that psychodynamic content can be increased by providing *Ss* with a moving stimulus (1).

This assumption was tested by using criteria of productivity and fantasy equally relevant to the Rorschach and PMS. Criteria of productivity were: (*a*) number of cards or sequences rejected, (*b*) number of human movement responses, and (*c*) number of animal movement responses. Total *R* was not used as a criterion since Rorschach and PMS response units are not strictly comparable. Structurally, PMS protocols more closely resemble TAT stories than Rorschach records. Three measures of fantasy production were obtained: (*a*) total number of unconscious fantasies, indicated by archaic, bizarre, or highly personalized content, further categorized into (*b*) static objects or fragments of personalized content, i.e., "womb," "broken arm," "dead person," and (*c*) descriptions of active elaborated processes, i.e., "embryo being born," "arms and legs falling off," "cancer inside my body eating away."

The sample consisted of 40 schizophrenics who were given both tests. Records were analyzed by the sign-rank test, showing results significant beyond the .01 level in favor of the PMS on the following criteria: rejection, human movement, animal movement. Total number of personalized fantasies did not differ significantly; however, number of "active elaborated processes" was significantly greater on the PMS (.01), and number of "static objects" was significantly greater on the Rorschach (.01). In the sample, 32.5% of *Ss* rejected more Rorschach cards than PMS sequences, 60% saw more human movement on the PMS, 75% saw more animal movement on the PMS, 62.5% had more personalized fantasies on the PMS, 82.5% had more static fantasies on the Rorschach, and 90% saw a greater number of elaborated fantasies on the PMS. Those *Ss* who rejected Rorschach cards responded to every PMS sequence.

Several hypotheses are suggested: (*a*) stimuli in motion lower the threshold for *M* type responses, (*b*) many times the Rorschach reflects the *S*'s affective dissociation and gives no clear evidence for the unconscious fantasies underlying his behavior, (*c*) for most of these *Ss* the PMS seems to give the necessary "push" towards elaborating fantasies and revealing affect which the clinician is then able to interpret in an unambiguous manner.

Brief Report.
Received May 11, 1956.

Reference

1. Lundin, William H. Projective movement sequences in the study of personality. *J. proj. Tech.*, 1954, **18**, 208–220.

---

[1] Based on a paper read at the 1956 meeting of the Midwest. Psychol. Ass., St. Louis, Mo.

[2] An extended report of this study may be obtained without charge from William H. Lundin, Psychology Dept., Chicago State Hospital, Chicago 34, Ill., or for a fee from the American Documentation Institute. Order Document No. 4972 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

A Continuous Performance Test of Brain Damage[1,2]

H. Enger Rosvold,[3] Allan F. Mirsky,[3] Irwin Sarason,[4]
Edwin D. Bransome, Jr.,[5] and Lloyd H. Beck[6]
Yale University

This paper presents the results of an investigation using a new instrument for the study of brain damage in human subjects. The design of the instrument, the Continuous Performance Test (CPT) was based on certain electroencephalographic evidence which suggested that brain-damaged individuals should show inferior ability as compared with non-brain-damaged individuals on tasks requiring sustained attention or alertness.

The waking EEGs of brain-damaged patients generally show either random bursts of hypersynchronous (high amplitude) activity intruding upon the normal activity of the brain, or a general hypersynchrony (3, 8). Hypersynchronous activity is also evident in the recording from the brain of a sleeping subject (3). If hypersynchrony is associated with reduced vigilance or attention, as suggested by its presence during sleep, then the hypersynchrony of the brain-damaged patients might also indicate reduced attention. For example, according to Hebb (1), hypersynchrony might interfere with the sequential firing of cell assemblies and thereby disrupt the process of attention. Other research strongly suggests that some such relationship does in fact exist. Thus, on the basis of a study of 75 prefrontal lobotomy cases, Levin states:

> Immediately following operation the patient usually lapsed into a drowsy-anergic state from which he could be temporarily aroused by stimuli, only to lapse once again into somnolence when the stimuli were removed. During this drowsy-anergic state EEGs showed diffuse slow activity and a good deal of baseline oscillations of $\frac{1}{2}$ to 2 per second anteriorly mixed with slow rhythms of 2 to 6 per second frequency. As the patient emerged from the drowsy stage the preoperative pattern usually asserted itself (2, p. 422).

The usual measures of attention or alertness such as the digit span and digit symbol substitution subtests of the Wechsler-Bellevue have not consistently showed decline following brain damage (4, 5, 6, 7). In the case of the patient who shows intermittent bursts of hypersynchronous activity, and hence only momentary lapses in attention, the test performance may be due to the fact that the *S* can to a great extent choose his own time to respond, and may reorganize his attention between momentary lapses. The lapses would then not affect his score to any measurable extent.

In the case of the individual who shows a generalized hypersynchrony of the electrical

---

[1] The authors would like to extend their appreciation to the following individuals and institutions for providing advice and/or subjects and testing facilities for this research: Mr. Samuel Greenhouse of the National Institute of Mental Health, Bethesda, Maryland; Dr. C. Edward Stull and the Southbury Training School, Southbury, Connecticut; Drs. Walter Landmesser and Russell Fuldner and the Newington Crippled Children's Hospital, Newington, Connecticut; Dr. Rhoades and the New Haven Cerebral Palsy Clinic, New Haven, Connecticut; Drs. Milton Senn, William German, Janice Stevens, Morris Wessell, Ethelyn Klatskin, and the staff of the several pediatric and neurological clinics of the Grace-New Haven Community Hospital, New Haven, Connecticut.

[2] Supported in part by grants to H. Enger Rosvold from the Veteran's Administration, Contract VA1001-M3222 and the National Science Foundation.

[3] Now at the National Institute of Mental Health, Bethesda, Maryland.

[4] Now at the West Haven V.A. Hospital, West Haven, Connecticut.

[5] Now at the College of Physicians and Surgeons, Columbia University, New York, N. Y.

[6] Now at the University of Michigan, Ann Arbor, Michigan.

activity of the brain, the classical tests might not reflect his deficiency because they do not require sustained attention over a sufficiently long period of time.

These considerations suggested that a test which would not allow the patient to choose his own time to respond, and which would require a high level of continuous attention over an appreciable interval of time might reflect a deficit that other procedures would miss. Accordingly, the CPT was designed to provide two attention tasks, labeled *X* and *AX*, with *AX* designed to be the more difficult. This paper presents data on the CPT performance of comparable groups of brain-damaged and non-brain-damaged individuals.

If the research demonstrates that there are indeed differences between brain-damaged and non-brain-damaged groups on this test, one of the questions which arises is whether or not the differences are large enough to be of diagnostic use. The data also provide a preliminary answer to this question.

Method

Subjects. No attempt was made in this study to select patients on the basis of either locus, extent, or type of brain damage. The term "brain-damaged" is here used to include any individual for whom there was medical evidence (surgical notes, neurological examination, EEG) of brain pathology. Three groups were studied, each consisting of a brain-damaged and a non-brain-damaged subgroup. The *Defective* group consisted of 72 institutionalized feebleminded *S*s, who were diagnosed as either of "organic etiology" (the *Brain-Damaged-Defective* subgroup) or of "familial or idiopathic etiology" (the *Non-Brain-Damaged-Defective* subgroup). The *Child* group comprised 45 children from a children's hospital and from various pediatric clinics who were being treated for brain disorders such as cerebral palsy (the *Brain-Damaged-Child* subgroup) or for non-nervous-system disorders (the *Non-Brain-Damaged-Child* subgroup). The third group, designated the Adult group, consisted of 50 adult *S*s. Half of these (the *Brain-Damaged-Adult* subgroup) were either epileptics being treated at a seizure clinic or brain surgery patients being seen at a neurosurgery clinic. The remaining 24 *S*s (the *Non-Brain-Damaged-Adult* subgroup) were either part of the general medical population of a hospital or were nonpatient *S*s.

The age and a measure of intelligence were obtained for all *S*s. In addition, the length of institutionalization was obtained for the Defective group. The *S*s in the Defective group were routinely tested on the Stanford-Binet upon admission to the training school; all other *S*s were tested for IQ just prior to presenting them with the CPT. The intelligence measure used for the Child group was the Stanford-Binet vocabulary, while the *S*s in the Adult group were tested on either the Full Scale Wechsler-Bellevue, the Verbal Scale of the Wechsler-Bellevue, or the 20-minute modification of the Otis Self-administering Test of Mental Ability, Form B. Approximately equal numbers of *S*s in each of the Adult subgroups were tested on each of the three measures of intelligence. The average ages and IQs of the several groups are presented in Table 1. The mean age of the Non-Brain-Damaged-Child subgroup was significantly higher than that of the Brain-Damaged-

Table 1

Mean Ages and IQs of the Subjects

Group	*N*	Mean age	*SD* age	Mean IQ	*SD* IQ
Defectives					
Brain-damaged	29	27.3	8.1	63.3	7.8
Non-brain-damaged	43	25.1	7.8	66.8	7.5
Children					
Brain-damaged	19	9.0**	3.0	101.9	28.2
Non-brain-damaged	26	12.3**	2.9	104.6	21.2
Adults					
Brain-damaged	25	31.7	12.1	90.5*	15.2
Non-brain-damaged	25	31.5	13.9	100.0*	15.0

* Indicates a significant difference ($p < .05$).
** Indicates a significant difference ($p < .01$).

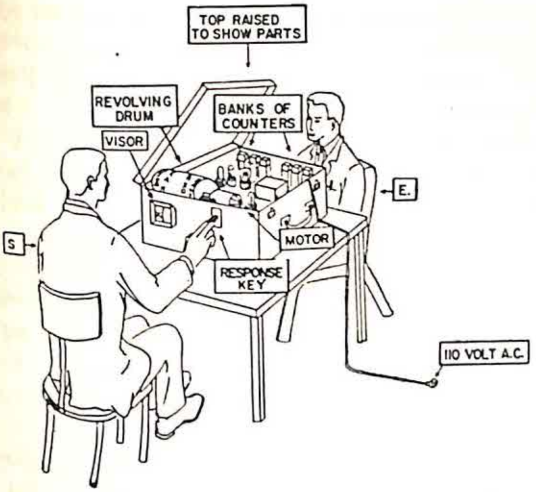


Fig. 1. Semischematic diagram showing the Continuous Performance Test apparatus, a subject and examiner in the positions they would occupy during a test.

Child subgroup ($p < .01$). Also, the IQ of the Non-Brain-Damaged-Adult subgroup was significantly higher than that of the Brain-Damaged-Adult subgroup ($p < .05$). There were no other significant differences between the various paired subgroups with respect to age or IQ. The mean length of institutionalization for the Brain-Damaged-Defective subgroup was 11.4 years, for the Non-Brain-Damaged-Defective subgroup, 9.3 years. This difference was not significant.

Apparatus. The testing instrument is illustrated in Fig. 1. It consisted essentially of a revolving drum on which two series, each of 31 letters, were mounted side by side.

One series served for the *X* task, the other for *AX* task. The drum revolved slowly, approximately twice a minute, in a boxlike case equipped with a visor through which the *S* was required to look. The *E* shifted the task from *X* to *AX* by changing the horizontal position of the visor. The letters were illuminated briefly one at a time. The illumination was provided by a bank of five 0.5-watt neon bulbs mounted inside the case, fired by means of the discharging of an 8-microfarad condenser. The *S* performed by pressing a response key when certain "correct" letters appeared. In the *X* task, every *X* was a correct letter; in the *AX* task, every *X* following an *A* was a correct letter. The letters came into view at approximately 0.92-sec. intervals. A response was scored correct if the response key was pressed within 0.69 seconds after the brief illumination of the correct letter. Following this interval, there was a period of approximately 0.23 seconds before the next letter was illuminated. Correct and incorrect responses were recorded automatically on a set of counters facing the back of the machine where *E* sat. The *X* task was always presented first and included eight *X*'s in the series of thirty-one letters. The responses occurring over the first two revolutions of the drum served as practice and were not counted in the *S*'s score. The *S*s were then given either a 5-minute test (10 revolutions) in which there were 80 possible correct *X* responses, or a 10-minute test (20 revolutions) in which there were 160 possible correct responses. Following a 2-minute rest period, the visor was moved by *E* so that the *AX* series was in view. There were six correct *AX* sequences among the thirty-one letters of the *AX* series. As with the *X* task, the responses occurring during the first two drum revolutions served as practice. Then, the *S*s received either a 5-minute test in which there were 60 possible correct *AX* responses, or a 10-minute test, in which there were 120 possible correct.

Procedure. No attempt was made to keep the instructions completely identical from *S* to *S;* the typical form, however, did not vary greatly from this content:

"Do you see this viewer? When the machine starts, you will see letters appearing one at a time. Your job is to press this key every time you see an *X*. Don't press it for any other letter and always press it when you see an *X*. Press the key down for a bit when you see an *X;* don't hit it too quickly. Try holding your eyes different distances from the viewer until you find the one that is most comfortable for you. Remember, press the key every time you see an *X*, but not for any other letter." Those *S*s who had difficulty in understanding were given further instruction of the same sort.

With the *X* task and the two-minute rest period completed, the *S* was instructed as follows:

"Your job in this next part is again to press the key when you see an *X*, but this time, only if the *X* follows right after an *A*. That is, when you see the series *A—X—* press the key on the *X*. Don't press the key for any other letter following an *A* except *X*, and always press it then. Remember now, when you see the letter *A*, get set; if an *X* comes right after it, press the key." Additional instruction was given to those who had difficulty in understanding.

Except for the children, all *S*s were tested for 10 minutes on each of the two tasks. With children, the time was reduced to 5 minutes on each task, and in addition, the instructions were modified to some extent, presenting the task as a game. For the children under the age of 10, pretesting was carried out to make certain that they could discriminate letters, and to see whether they could follow the instructions by requiring answers to two special test sheets, one containing *X*s among other letters, the other, *A*s and *X*s among other letters. Those who failed this pretest were not tested further. No *S* was tested who could not perform the minimal motor tasks of holding his head up unsupported and pressing the response key. Since the testing was conducted at many different clinics and hospitals, it was not feasible to control exactly the amount of illumination in the testing room. The ordinary lighting of the room was used without any modification; care was taken however, to insure that there was no glare on the visor of the testing apparatus.

Reliability. Two measures of the reliability of the CPT were made, using *S*s of the Defective group. Odd-even reliability was determined from the scores of 21 non-brain-damaged *S*s, by comparing the subtotal of correct responses on the 10 even revolutions of the drum with the subtotal on the 10 odd revolutions. The Pearson r for the *X* task was 0.88, for the *AX* task 0.86. To determine test-retest reliability, 21 brain-damaged and 22 non-brain-damaged *S*s were retested from four to seven weeks after the initial testing. The test-retest r for the number of correct responses of the brain-damaged group on *X* was 0.90, for *AX*, 0.79. The test-retest r for the number of correct responses of the non-brain-damaged group on *X* was 0.88, for *AX*, 0.74.

CPT scores. The *X* and *AX* responses of each *S* were scored in two ways. The absolute percentage correct was defined as the number of correct responses the *S* made divided by the number of correct responses (× 100) possible in the time allowed. The relative percentage correct was defined as the number of correct responses (× 100) the *S* made divided by his number of attempts (both correct and incorrect) during the time allowed. In the event that the two scores differed markedly, depending on the *S*'s pattern of responding, the two measures provided a more complete description of his performance than would either of the two taken singly.

Results

CPT scores. Table 2 presents the means and standard deviations of the absolute percentage of possible correct responses on the *X* and *AX* tasks for the six subgroups. Since there was a significant age difference between the Child subgroups and a significant IQ difference between the Adult subgroups (Table 1), the differences between the paired subgroup means on *X* and *AX* in these two groups were evaluated by means of an analysis of covariance. The means of the Child group in Table 2 have been adjusted for regression on age and the means of the Adult group have been adjusted for regression on IQ. The covariance procedure was corrected for heterogeneity of variance between the subgroups when the analysis required it. The mean differences in the Defective group were evaluated by means of the t test.

The differences between the subgroup means were highly significant on both the *X* and

Table 2

Adjusted Means and Standard Deviations for Child and Adult Groups, Means and Standard Deviations for Defective Groups, *X* and *AX* Absolute Percentage Correct Measures

Group	Brain-damaged			Non-brain-damaged			
	N	Adj. mean	*SD*	*N*	Adj. mean	*SD*	p diff.
Children							
X	19	41.61	28.39	26	82.57	14.65	<.0005
AX	19	46.46	28.20	26	79.41	17.44	<.0005
Adults							
X	25	78.53	28.26	25	88.00	17.21	>.30
AX	23	61.40	26.66	23	78.77	15.33	<.01
Defectives							
X	29	46.44	32.02	43	67.86	25.62	<.01
AX	29	32.30	25.79	43	54.74	20.02	<.0005

Table 3

Significance Levels of Differences Between X and AX Scores, for Relative and Absolute Measures of Performance, All Groups

	Brain-damaged		Non-brain-damaged	
Group	Absolute	Relative	Absolute	Relative
Children	$>.70$	$>.30$	$>.50$	$>.70$
Adults	$<.001$	$<.01$	$<.01$	$<.05$
Defectives	$<.01$	$<.001$	$<.001$	$<.001$

Table 4

Significance Levels of Differences Between Adjusted AX Mean Scores of Brain-Damaged and Non-Brain-Damaged Subgroups, All Groups

	Performance measure	
Group	Absolute	Relative
Children	$>.20$	$<.05$
Adults	$<.01$	$>.05$
Defectives	$<.05$	$<.05$

AX tasks in the Defective group ($p < .01$ and $p < .005$, respectively) and also in the Child group ($p < .0005$ in each case). In the Adult group, the brain-damaged subgroup was significantly inferior to its controls on AX ($p < .01$), but not on X ($p > .30$).

When the *relative* percentage scores of the paired Child and Adult subgroups were compared by means of an analysis of covariance, and those of the Defective group by means of the t test, the paired subgroup differences were all in the same direction and of approximately the same magnitude as those reported in Table 2. The levels of significance were in no case less than the corresponding levels reported in Table 2. The only difference was that the *relative* scores on the X task discriminated significantly ($p < .05$) between the Adult subgroups, whereas the absolute scores on the X task (Table 2) did not.

In order to determine if the AX task was in fact more difficult than the X task, the mean X score was compared with the mean AX score for each of the six subgroups. Table 3 presents the significance levels from the t test of the differences between X and AX scores for both the relative and absolute measures of performance in all six subgroups. Each of the Defective and Adult subgroups performed significantly worse (at $p < .05$ or less) on the AX than on the X task. With the children, none of the differences between X and AX was significant. To determine if the brain-damaged subgroups had relatively more difficulty on AX (i.e., independent of their performance on X) than the non-brain-damaged subgroups, an analysis of covariance was run on the AX scores of the paired brain-damaged and non-brain-damaged subgroups with X scores as the covariate. Table 4, which presents the results of the analysis, shows that in terms of the absolute measures

Table 5

Percentages of Correct Identification of Diagnostic Category by CPT Scores, All Groups

	Group					
	Defectives		Children		Adults	
Subgroup	Brain-damaged	Non-brain-damaged	Brain-damaged	Non-brain-damaged	Brain-damaged	Non-brain-damaged
N	29	43	19	26	25	25
$X-R$	75.9***	67.4**	84.2***	76.9***	68.0*	68.0*
$X-A$	62.0	58.1	89.5***	76.9***	64.0	64.0
$AX-R$	72.4**	65.1*	84.2***	76.9***	66.7*	66.7*
$AX-A$	75.9***	67.4**	84.2***	76.9***	65.0†	65.0†

* $p < .05$.
** $p < .01$.
*** $p < .005$.
† Based on 23 subjects.
Note.—p-values refer to the probability that the percentage differs from chance (50%). Nonstarred percentages do not differ significantly from chance.

of performance, the Brain-Damaged-Adult and -Defective subgroups had significantly greater difficulty on the *AX* task than their respective nondamaged subgroups, even when the paired subgroups were adjusted to equal levels of performance on the *X* task. Similarly, in terms of the relative measures of performance, the Brain-Damaged-Child and -Defective subgroups had significantly greater difficulty on *AX* than their respective nondamaged subgroups.

Correct identification of subgroup. Table 5 presents the percentages of correct identification of diagnostic category of all the subgroups on all measures yielded by the CPT. These were determined for each subgroup on each CPT measure from the over-all median of scores of a group on that measure. For example, considering the *X* scores in the Defective group, the percentage of brain-damaged *Ss* scoring *below* the combined Defective group median defines the percentage of correct identification of brain-damaged defectives on the *X* task. Similarly, the percent-

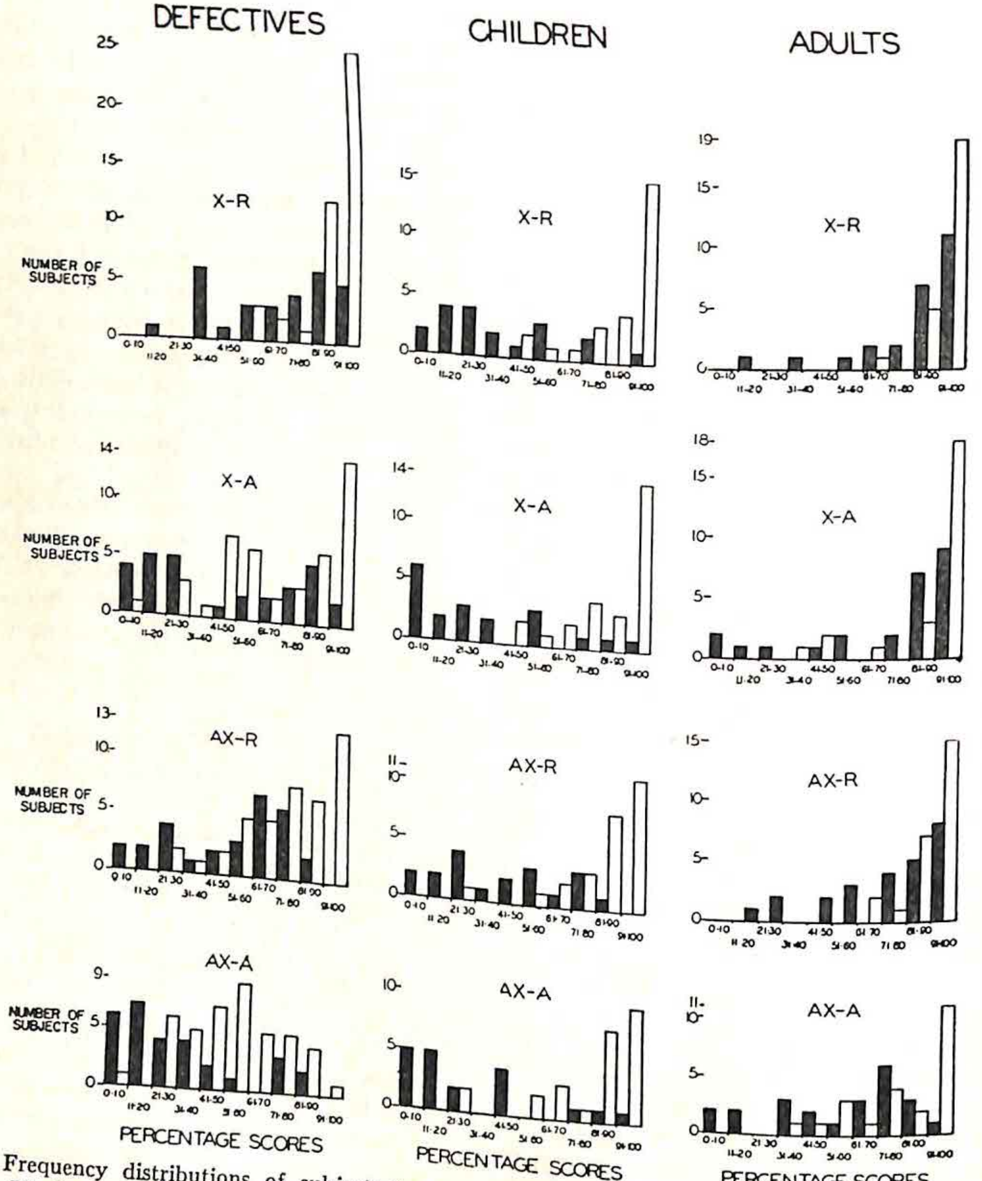


Fig. 2. Frequency distributions of subjects by percentage correct score intervals, all groups, all CPT measures. Black columns represent frequencies of brain-damaged subjects, white columns, frequencies of non-brain-damaged controls.

age of non-brain-damaged defectives scoring *above* the over-all median defines the percentage of correct identification of non-brain-damaged defectives on the *X* task. These percentages were computed to provide an estimate of the potential usefulness of the several scores of the CPT as diagnostic indicators in various clinical groups. In Table 5, *X-R* and *AX-R* refer to the relative measures of performance on these tasks; *X-A* and *AX-A* refer to the absolute measures of performance. The relative measures provided more significant percentages of correct identification than did the absolute measures; all of the six *X-R* and all of the six *AX-R* subgroup percentages were significant at $p < .05$. The Child group had more significant percentages of correct identification than any other (eight of eight) all of which were significant at $p < .005$. There were equal numbers of significant percentages of correct identification among the non-brain-damaged and brain-damaged subgroups; however, these percentages tended to be higher among the brain-damaged subgroups.

Figure 2 presents for each group and CPT measure, the distribution by percentage score of brain-damaged and non-brain-damaged *S*s (uncorrected for age or IQ differences between subgroups). Figure 2 emphasizes in general the great range in scores achieved by brain-damaged *S*s as compared with non-brain-damaged *S*s. The scores of the non-brain-damaged *S*s are usually concentrated near the upper end of the distribution; in fact, the modal score for non-brain-damaged *S*s is in the 91 to 100% interval in every case except for the *AX-A* measure in the Defective group. In contrast with this, the scores of the brain-damaged *S*s span the entire distribution, or the entire distribution excepting one score interval, in every case.

Discussion

The results of this study indicate that brain-damaged individuals perform poorly relative to non-brain-damaged controls on a task requiring continuous attention. The results also indicate that on a task designed to require even more sustained attention the brain-damaged individuals perform even more poorly. The score distributions of the various subgroups (Figure 2) and the percentages of individuals correctly identified by diagnostic category (Table 5) suggest that the CPT is sufficiently sensitive to the effects of brain damage so that it might prove useful clinically. The data of Table 5 suggest that the relative measures are more discriminating than the absolute measures, that the test is somewhat more reliable in reflecting the presence than the absence of brain damage and that it is most discriminating when applied to the children of this study.

These findings are of course preliminary. Use of the CPT in the clinic must await standardization on a variety of diagnostic groups, to ascertain whether brain pathology only and which types of brain pathology lead to deficient performance. For example, some data which have not been reported here indicate that schizophrenics perform no differently from normals of comparable age and IQ. More information concerning other groups is required. Standardization on a variety of diagnostic groups is also necessary in order to have a reliable diagnostic score for each subgroup. Thus, in the present study, cutoff points which included most brain-damaged children missed many brain-damaged adults (see Figure 2).

Diagnostic considerations aside, the question remains as to whether the impaired performance of the brain-damaged *S*s on the CPT is in fact due to an impairment of attention, whether in the form of momentary lapses, or of a lowering of a general level. Clearly, the available information does not rule out alternate interpretations of the deficit. To make certain that it is attention that is impaired would require more extensive information about the *S*'s test behavior than the present instrument provides. It is clear, furthermore, that the available information does not permit the conclusion that the impaired CPT performance is indeed related to hypersynchrony as evident in the EEG. Nonetheless, the results of this study do lend support to these hypotheses suggesting that it might now be profitable to make a more direct test of the original formulation. This test would be provided by simultaneously measuring CPT performance and brain activity to determine whether errors on the CPT are coincident in time with hypersynchronous patterns on the EEG and whether there is a correlation between degree of deficit on the CPT and degree of hypersynchrony.

Summary

1. The Continuous Performance Test, a procedure for the detection and study of brain damage in humans, is described.

2. Three groups of *S*s, each including a brain-damaged and a non-brain-damaged subgroup, were tested on this procedure.

3. The brain-damaged subgroups were significantly inferior to their non-brain-damaged controls on the measures yielded by the CPT, and these differences were increased when the difficulty of the task was increased.

4. The CPT is sufficiently reliable and yields sufficiently large differences between subgroups to suggest that it might ultimately prove useful as a clinical instrument for the diagnosis of brain damage.

5. An interpretation of the inferior performance of the brain-damaged *S*s was offered in terms of impairment in attention or alertness and suggestions were made about future research relating cerebral events and CPT performance.

Received January 13, 1956.

References

1. Hebb, D. O. *The organization of behavior.* New York: Wiley, 1949.
2. Levin, S., Greenblatt, M., Healey, M. H., & Solomon, H. C. Electroencephalographic and clinical effect of prefrontal lobotomy, with consideration of postlobotomy convulsive seizures. In M. Greenblatt, R. Arnot, & H. C. Solomon (Eds.), *Studies in lobotomy.* New York: Grune & Stratton, 1950. Pp. 400–427.
3. Lindsley, D. B. Electroencephalography. In J. McV. Hunt (Ed.), *Personality and the behavior disorders.* New York: Ronald, 1944. Pp. 1033–1103.
4. Mettler, F. A. (Ed.) *Psychosurgical problems.* New York: Blakiston, 1952.
5. Mettler, F. A. (Ed.) *Selective partial ablation of the frontal cortex.* New York: Hoeber, 1949.
6. Partridge, M. *Pre-frontal leucotomy.* Springfield, Ill.: Charles C Thomas, 1950.
7. Petrie, A. *Personality and the frontal lobes.* New York: Blakiston, 1952.
8. Schwab, R. S. *Electroencephalography in clinical practice.* Philadelphia: Saunders, 1951.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Stimulus Structure as a Determinant of the Perceptual Response[1]

M. Powell Lawton

Veterans Administration Hospital, Providence, R. I.

The role of formal stimulus characteristics is acknowledged to be of importance in the interpretation of projective test data. Clarification of the meaning of projective test responses has been sought through study of stimulus qualities in a number of areas (1, 2, 6, 8, 12, 14, 15).

With respect to the Rorschach test, the role of form in the production of a response has received surprisingly little attention. The concept of the popular response implies that formal blot characteristics of certain areas of the Rorschach tend to call forth a particular response. Ranzoni, Grant, and Ives (10) obtained a set of norms for an adolescent population specifying which cards were most likely to elicit responses using a particular location, determinant, and content. Their study was not concerned primarily with form, however, and made no attempt to vary the forms of the Rorschach stimuli or to place them in an order of complexity. Meer's study (9) established a tentative order of organizational difficulty among cards, but did not attempt to specify the factors believed to be responsible for stimulus ease or difficulty. Stein's (11) tachistoscopic exposure of the cards demonstrated that perception was less adequate under impoverished stimulus conditions but again, the stimulus condition as a whole was varied, with no attempt to focus on form as the variable for study.

The present study investigated by means of a non-Rorschach perceptual task the general hypothesis that stimulus ambiguity is an important determinant of perceptual adequacy. Degree of maladjustment was a second variable studied in line with Forer's (5) suggestion that degree of personality integration is a determining factor in the differential adequacy with which an individual perceives in situations of varying ambiguity.

It was hoped that variation of the form of the stimulus, isolated from as many other stimulus characteristics as possible, would provide a general basis upon which to interpret individual projective records, though the limits of applicability to any specific test are fully realized. As in Stein's (11) study, response characteristics were evaluated in terms of perceptual adequacy, defined so as to include both formal and content characteristics of the response. The specific hypothesis was that both increasing stimulus ambiguity and increasing degree of pathology would result in decreased perceptual adequacy.

Procedure

In order to limit investigation as much as possible to form alone, sixty figures were randomly cut from black paper and mounted on white cards, leaving only the blackness of the figure as a confounding factor. Seventeen psychologists rated each figure as most, least, or moderately structured, using as their criterion of structuredness the relative ease with which each figure suggested a percept. From these judgments five figures were clearly chosen as the most structured (Structure I), five as moderately structured (Structure II), and five as least structured (Structure III). The 15 figures were arranged in random order for standardized presentation to the subjects. This "subjective consensus" method of defining stimulus ambiguity, though it has the disadvantage of risking bias through the use of only psychologists as judges, affords a definition of ambiguity in psychological, rather than physical, terms. The psychological definition was deemed more appropriate to the present study.

---

[1] Adapted from a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Faculty of Philosophy, Columbia University, 1952. The author wishes to thank Dr. E. J. Shoben, Jr., Dr. Helen Walker, and Dr. Laurance Shaffer, members of his committee, for their help, and also Dr. Bernard Locke, Chief Psychologist, New York VA Hospital, for making possible the gathering of the data.

Thirty *Ss* were chosen from each of three nosological groups: neurotic, latent schizophrenic, and hospitalized schizophrenic, as diagnosed either by their outpatient therapists or ward physicians. By using this ordinal scale of degree of maladjustment, it was hoped that unreliability in ratings of maladjustment could be avoided. The three groups were matched for age and education. Analysis of variance and Bartlett's test showed no reason to suspect that either means or variances differed significantly. The mean age of the entire group was 29.8 years, mean educational level 11.4 years.

The cards were presented in standard position and *Ss* were instructed to tell the examiner what the whole figure could look like. Turning and detail responses were discouraged, though accepted if *S* could not respond in the requested manner. Only the first response to each figure was used. Through judicious urging only five rejections out of a possible 1,350 responses occurred. The *S* was asked to describe each response further. This was the examiner's only question, other than clarification of the areas used.

Table 1

Number of Patients Receiving All "*1*" Ratings on Form Level

Group	Stimulus Ambiguity I	II	III
Neurotics	28	19	14
Latent schizophrenics	25	13	8
Hospitalized schizophrenics	17	3	1

The following categories were employed in evaluating the perceptual quality of the responses:

Form Level: a four-point rating scale measuring judged goodness of fit of percept to stimulus. Scale points were named *1.* adequate form, *2.* vague form, *3.* poor form, and *4.* very poor form.

Structural Peculiarities: (*a*) *Fabulation* was a response which added a great deal more to the percept than could reasonably be assumed from the stimulus ("Miss Olsen's profile," "a person who died of pneumonia.") (*b*) *Incongruity* was a percept whose parts do not belong together ("a person with a meat cleaver attached to his head," "a rabbit—here's the opening in his shell.")

Sex Response: Any overt genital or anal response.

Other content categories were scored, but occurred infrequently and could not be scored reliably. A Total Pathology rating scale was also used, but because of its high correlation with the Form Level scale, added little further information.

Each response was typed and coded individually and the ninety responses to each figure were scored consecutively by the writer. One-third of the total responses were submitted for a reliability check to two other

Table 2

Number of Patients Without Structural Peculiarities, With Cochran's Q Test and Combined Probability Measure

Group	Stimulus ambiguity I	II	III	Q	p	Order probability	Combined probability
Neurotics	29	28	26	2.36	.308	1/6	.114
Latent schizophrenics	21	19	15	3.10	.214	1/6	.080
Hospitalized schizophrenics	20	19	18	.48	.794	1/6	.295
All groups	70	66	59	5.03	.082	1/6	.031

psychologists.[2] Through a method described in detail elsewhere (7), scoring reliability was determined and was deemed good enough to permit using the writer's scoring in the evaluation of the data.

Results

Table 1 shows the number of patients all of whose responses received Form Level scores of *"1,"* by clinical group and by stimulus structure. The number of patients receiving all *"1"* ratings decreased regularly both as stimuli became more ambiguous and as pathology increased. A nonparametric significance test was performed as follows. For nine cells there are 9! possible arrangements, and of these 9! arrangements only 42 can be found such that in every row the frequencies decrease from left to right and in every column decrease from top to bottom. Accordingly, the probability of obtaining by chance alone data showing this regular decrease in Form Level with increasingly ambiguity and with increasing degree of illness is 42/9!, or .0001156.

Table 1 suggested the possibility that the rate of decline in Form Level as stimuli become more ambiguous might be greater as psychopathology increases. This supposition was tested in the following manner. A rate of decline score was obtained for each *S* by subtracting the sum of his Form Level scores for Structure I from the sum of his Form Level scores for Structure III and dividing this difference by the sum of Structure I and Structure III Form Level sums. The mean rate of decline score for neurotics was .0759, for latent schizophrenics .1223, and for hospitalized schizophrenics .1935. An analysis of variance resulted in an *F* value of 4.95, significant at the .01 level for 2 and 87 degrees of freedom.

Table 3

Number of Patients Who Give No Structural Peculiarity

Group	Number*
Neurotics	25
Latent schizophrenics	11
Hospitalized schizophrenics	12

* Chi square = 7.61; $p < .05$.

The incidence of Structural Peculiarities is shown in Table 2. Columns I, II, and III indicate the number of patients who did not respond with a Structural Peculiarity to any one of the five figures of each structural level. In order to test the hypothesis that incidence of Structural Peculiarities is a function of stimulus ambiguity, Cochran's *Q* test (3), a variant of chi square that does not require independence of observations, was used. None of the probabilities (Column six, Table 2) associated with *Q* values reaches statistical significance, but when account is taken of the fact that the predicted order across rows was obtained in each case (13), the probability value for the entire sample is seen to be .031. The usual chi-square test applied to the relationship between degree of maladjustment and incidence of Structural Peculiarities (Table 3) gave a chi-square value of 7.61, associated with a probability value less than .05. Inspection reveals that the significance of this measure is derived largely from the relative absence of Structural Peculiarities in the neurotic group.

Table 4 indicates the number of patients showing no Sex responses to cards of different stimulus structures. Again using Cochran's

Table 4

Number of Patients Without Sex Responses, With Cochran's *Q* Test and Combined Probability Measure

	Stimulus ambiguity						
	I	II	III	*Q*	*p*	Order probability	Combined probability
All groups	86	84	81	3.80	.142	1/6	.053

---

[2] Dr. Gerald Bauman and Dr. Melvin Kornreich most kindly acted as judges.

Table 5

Number of Patients Who Give No Sex Response

Group	Number*
Neurotics	29
Latent schizophrenics	22
Hospitalized schizophrenics	25

* Chi square = .98; $p > .05$.

Q test and combining probabilities, the association between stimulus structure and incidence of Sex responses is seen to be of borderline significance ($p = .053$).

Table 5 indicates the number of patients giving no Sex responses. The chi-square value of .98 is associated with a probability greater than .05. Accordingly the hypothesis that incidence of Sex responses is a function of degree of maladjustment must be rejected.

Discussion

Perceptual distortion as measured by form accuracy is thus seen to be a function of both stimulus ambiguity and degree of maladjustment. On the other hand, perceptual distortion as measured by indices relating to content and the organization of content show a highly tenuous, if at all significant, relationship to stimulus ambiguity and to degree of maladjustment.

It appears clear, then, that it is important to consider stimulus characteristics in the interpretation of a projective response. In the Rorschach, for example, a response judged to be of poor form quality would have different meaning depending upon whether it occurred to a highly structured area or to a relatively amorphous area. As indicated in Table 1, a poor form response to the most ambiguous stimuli would give less information about the presence of a schizophrenic process in an individual than would a poor form response to a highly structured stimulus: the original data show that four neurotics gave ten definitely poor form responses to the less well-structured cards, but none to the best structured. It can be inferred that any response of less than adequate form level to a popular area of the Rorschach cards should make the interpreter suspicious of a severe disturbance, whereas an accumulation of $F-$ responses on less well-structured areas would be required to give the same impression. The present results would tend to confirm Meer's (9) supposition that an important factor in determining Rorschach card difficulty is stimulus ambiguity. A problem for further research might involve the development of a more discriminating Rorschach $F+\%$ than is presently used by weighting more heavily those $F-$ responses that occur to highly structured areas.

On the other hand, the equivocal results obtained in the attempt to relate Structural Peculiarities and Sex responses to stimulus structure make interpretation somewhat difficult. It may be that there is actually a greater contribution of internal factors than of stimulus factors in the production of these content-related distortions, though the present study does not answer this question. This supposition is reasonable if one considers that standards of what is appropriate in matters of content are largely internal, in distinction to the Form Level dimension, where the model against which the appropriateness of the *S*'s association is tested is physically present in the stimulus.

The differential rates of decline in Form Level with increasing ambiguity among the three clinical groups offers experimental verification of the observation that the more maladjusted individuals "go to pieces" more easily than the less maladjusted. Not only is there a difference in adaptive ability, but this difference becomes wider as the external situation becomes more demanding, in the sense of forcing the individual to provide structure to an ambiguous field from within himself. Such a finding would suggest that the best differentiation among groups according to degree of maladjustment would come from rating the Form Level of the more ambiguous cards. This supposition was tested by Feldman, Gursslin, Kaplan, and Sharlock (4) who found that an $F+\%$ which excluded popular responses from consideration was more discriminating between normals and neurotics than Beck's or Hertz's $F+\%$.

Summary

As a means of exploring the Form Level dimension in projective perception, a method

was devised to test the hypothesis that both increasing stimulus ambiguity and increasing personality maladjustment would have an adverse effect upon perceptual adequacy. Black paper cutouts of three ranked degrees of ambiguity were presented in Rorschach fashion to three groups of patients representing three degrees of maladjustment. It was found that both greater stimulus ambiguity and greater personality maladjustment were associated with lesser perceptual adequacy, as measured by Form Level. Perceptual adequacy as measured by several indices involving response content and organization of content was seen to have an equivocal relationship to stimulus ambiguity and to personality maladjustment. It was concluded that stimulus characteristics are often important determinants of projective responses and that consideration of form ambiguity is necessary to the best interpretation of projective records.

Received November 21, 1955.

References

1. Allen, R. M. The influence of color in the Rorschach test on reaction time in a normal population. *J. proj. Tech.*, 1951, **15**, 481–485.
2. Bijou, S., & Kenney, D. T. The ambiguity values of TAT cards. *J. consult. Psychol.*, 1951, **15**, 203–209.
3. Cochran, W. The comparison of percentages in matched samples. *Biometrika*, 1950, **37**, 256–266.
4. Feldman, M. J., Gursslin, C., Kaplan, M. L., & Sharlock, N. A preliminary study to develop a more discriminating $F+$ ratio. *J. clin. Psychol.*, 1954, **10**, 47–51.
5. Forer, B. The latency of latent schizophrenia. *J. proj. Tech.*, 1950, **14**, 297–302.
6. Kenney, D. T., & Bijou, S. Ambiguity of pictures and extent of personality factors in fantasy responses. *J. consult. Psychol.*, 1953, **17**, 283–288.
7. Lawton, M. P. Stimulus ambiguity and personality maladjustment as related to perceptual adequacy. Unpublished doctor's dissertation. Columbia Univer., 1952.
8. Lazarus, R. The influence of color on the protocol of the Rorschach test. *J. abnorm. soc. Psychol.*, 1949, **44**, 506–516.
9. Meer, B. The relative difficulty of the Rorschach cards. *J. proj. Tech.*, 1955, **19**, 43–53.
10. Ranzoni, J., Grant, M., & Ives, V. Rorschach "card-pull" in a normal adolescent group. *J. proj. Tech.*, 1950, **14**, 107–133.
11. Stein, M. Personality factors involved in the temporal development of Rorschach responses. *J. proj. Tech.*, 1949, **13**, 353–414.
12. Wallen, R. The nature of color shock. *J. abnorm. soc. Psychol.*, 1948, **43**, 346–356.
13. Wallis, W. Compounding probabilities from independent significance tests. *Econometrika*, 1942, **10**, 229–248.
14. Weisskopf, Edith. A transcendence index as a proposed measure in the TAT. *J. Psychol.*, 1950, **29**, 379–390.
15. Weisskopf, Edith. An experimental study of the effect of brightness and ambiguity on projection in the TAT. *J. Psychol.*, 1950, **29**, 407–416.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Manifest Anxiety and the Palmar Perspiration Index[1]

Allen D. Calvin, F. J. McGuigan, Sybil Tyrrell,
and Marshall Soyars
Hollins College

During the last few years the Taylor A scale has become an increasingly important measure of manifest anxiety. During the same period Mowrer and others from the University of Illinois have been developing the Palmar Perspiration Index (PPI) as a measure of psychological tension. The present study is an attempt to investigate the relationship between the A scale and the PPI.

The A scale and the higher form of the Otis were given to 54 female undergraduate students. In order to create a stress situation the *S*s were called in for individual experimental sessions and given the Wechsler digit span. Immediately after completing the digit span a PPI was obtained. After approximately three months *S*s were asked to fill out a six-point self-rating anxiety scale.

Product-moment correlation coefficients were obtained between the following: A scale and PPI − .01; A scale and digit span − .05; digit span and PPI − .16; Otis and A scale − .02; Otis and PPI .22. None of these was significant. A biserial correlation of .35 ($p < .05$) was obtained between self-ratings and A scale scores, and a biserial of .17 (not significant) was obtained between self-ratings and the PPI.

The failure to find a correlation between A-scale scores and PPI scores was unexpected since at first glance both seem to be measuring somewhat the same variable. However Beam (1) in a study published after the completion of our experiment, reports that he too found no significant relationship between the PPI and A scale. He points out that his Taylor scores fell within the middle range and were fairly homogeneous. Our A-scale scores had a mean of 16.9 with an *SD* of 6.9, indicating considerable scatter within our group. It seems highly unlikely that our failure to obtain a significant relationship between A scale and PPI scores can be ascribed to too much homogeneity within our group.

On the basis of our findings and other studies of anxiety the following hypothesis is proposed: The A scale and the PPI measure two different facets of behavior. Where anxiety is of a temporary "ego-involved" situational nature, the appropriate measure of drive is the PPI. Conversely, where anxiety is a stable permanent personality characteristic of the individual the appropriate measure of drive is the A scale.

Brief Report.
Received April 26, 1956.

Reference

1. Beam, J. C. Serial learning and conditioning under real-life stress. *J. abnorm. soc. Psychol.*, 1955, **51**, 543–551.

---

[1] An extended report of this study may be obtained without charge from Allen D. Calvin, Hollins College, Va., or for a fee from the American Documentation Institute. Order Document No. 4849 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Effective Intelligence as Measured by an Unstructured Sentence-Completion Technique[1]

George E. Copple
Vanderbilt University

Numerous varieties of test material and test procedure have been used in assessing the intellectual functioning of children. Almost all these procedures have involved rather strict structuring, and this structuring has ordinarily been considered a merit of the procedure. Perhaps because of this emphasis on easily scored, "right or wrong" answers, the possibilities of using a sentence-completion technique to evaluate intelligence have been largely overlooked. The present study will describe a sentence-completion technique which has been developed as a measure of "effective intelligence" and will present data on its usefulness.

Earlier Uses of Sentence Completions

Beginning with Ebbinghaus (1), and best known from its inclusion on the Stanford-Binet (2), one type of sentence completion has long been used as an indicator of intelligence. In this type, most of the sentence is given, with only a connective word omitted. Acceptable choices for this omitted word are quite limited and an answer is either "right" or "wrong." M. R. Trabue (3) used a similar method in a large-scale study. He called his test "language scales" and it is obvious that knowledge of word meanings and language usage played the greatest part in determining a successful performance.

The other, and more recent, use of sentence completions has been to evaluate personality or adjustment. Every survey of projective techniques is likely to have a section or a chapter on sentence-completion techniques. These techniques are usually for adults, they consist of "loaded" stems intended to tap specific areas of adjustment ("My greatest weakness is . . ." or "I think that most mothers . . ."), and the responses are evaluated for content, particularly dynamic content. The level of ability shown by the completions and the range and depth of interests are ignored.

The use of the sentence-completion technique to be described in the present paper is different from these other uses reported in the literature in general purpose, technique of administration, and rationale.

Rationale

Most intelligence-test items demand little that is creative or even integrative of the person being tested. They provide a task or a problem that has a narrow range of correct answers. They provide a great deal of structuring and often call for a single word or phrase or brief explanation. Typical items are: "What is the thing to do if you are on your way to school and see that you are in danger of being late?" "What is the color of rubies?" "How are a coat and a dress alike or the same?" "*Sky* is to *blue* as *grass* is to ———." Undoubtedly items of this kind give some valid indication of a subject's intelligence. But when a test is composed exclusively or mostly of such items, some important aspects of intelligence are left out. The individual who can readily place Brazil in South America, Egypt in Africa, or India in Asia, may not be able to respond successfully if asked instead to name one country in each of these continents. With still less structuring, as for example in responding to the instruction to give the best sentence completion he can to the stimulus "In other countries . . .," he may give us something inept and vague such as "things are different" or "they are foreigners." One limitation of virtually all intelligence tests is just this failure to let the subject do enough—the failure to let him show us what he can do and does do when the stage is not too completely set.

---

[1] This article represents a revision of a paper presented at a meeting of the Southern Society for Philosophy and Psychology in Atlanta, Georgia, on April 16, 1954.

An example may be useful. Here are two nine-year-old boys who are completing the stimulus-stem "A piece of string. . . ." One completes it "is long." The other supplies "can be used to make a bow and arrow." This latter is obviously a far superior answer. It is most unlikely, however, that the first boy was ignorant of the fact which the second boy's completion utilized. On a more conventional test item calling for this information, such as, "A bow is made of ——— and ———," both boys would probably be able to answer correctly. But does not the second boy's method of completing the sentence show something in addition to the possession of information? Does not this "something" seem to be basic to many types of school and nonschool performance that we label "intelligence"?

Administration of the Test

The Copple Sentence Completion Test is composed of 60 items, only 30 of which are given to any one subject. Level A, consisting of items 1–30, is for children 5–8 inclusive. Level B, consisting of items 16–45, is for children 9–12 inclusive. Level C, with items 31–60, is for children 13 and older, and for adults. Items at the various levels differ somewhat in difficulty, but this is not very easy to insure and is of minor importance. The scoring standards differ for the different levels of the test, so that even if an intrinsically easy item appears on the higher level test, it is more strictly scored and a better performance is required for a comparable rating.

Administration is always individual, with oral responses by the subject, which are recorded by the examiner. The procedure enables young children or illiterates to respond as fully as they are able, without penalizing or limiting them in any way. It permits people in all age groups to use natural conversational completions without regard to avoiding words they cannot spell. At the other extreme, it discourages "literary" completions. Experience seems to show that it is unnecessary to limit the time a subject can use in framing his responses. As a rule, the longer the delay, the poorer the response (except for extremely brief latencies).

In introducing the test, the subject is told that he will be given some words and that he is to add some more words so as to produce the very best sentence he can make. A sample stem is given the subject and two ways of completing it are presented, one being designated as "better" than the other. For example, the stem might be "The house. . . ." One way to complete it, the subject is told, would be to say "is large." Another, and a better way, would be to say "stood in the bend of the road." Nothing is said about *why* the latter is better, and subjects as young as five years accept this as a self-evident superiority.

With this brief introduction, the test is begun. It requires about 15 minutes to complete. It is rare, even at age five, to have a child say "I don't know" to even a single item. Few persons seem to experience any sense of failure in completing the sentences. The intrinsic interest is high and no matter at what level the individual is performing, he is likely to consider that he is doing well. This feature, of course, is a particular virtue when a whole battery of tests is to be given or when a subject is to be worked with therapeutically.

Scoring the Responses

Each response is scored 0, 1, 2, or 3 in accordance with prepared scoring standards. These standards reflect such considerations as relevance, complexity, descriptiveness, and range or scope. Numerous examples have been provided for each score value on each item. Judgment is still necessary, but in numerous studies the agreement in scoring has been high ($r = .95$ or above). For example, in one study using 24 protocols and with 10 psychology graduate students as scorers, none of whom had had any previous experience with the test, the median interscorer reliability was .97.

The split-half reliabilities of the test itself have been uniformly high. For the three studies reported below, these reliabilities (corrected by the Spearman-Brown formula) were respectively .94, .95, and .93. No test-retest data have been gathered.

Measuring "Effective Intelligence" (Validity)

The most valuable criterion for a test of "effective intelligence" would be the judg-

Table 1

CSCT Scores for Three Age Groups

Age Group	*N*	Mean Test Score	Sigma	*CR*
9–6	29	40.76	8.98	2.27 (.02 level)
10–6	17	47.71	10.14	1.66 (.05 level)
11–6	20	53.85	11.73	

Table 2

Correlations Between Age and Scores on Three Different Tests

Test	Sample A (*N*=33)	Sample A′ (A plus 17 additional)
CSCT (Form B)	.584	.577
Binet (Form L)	.621	
Goodenough DAM	.294	.400

ments of objective, well-trained observers on the actual behavioral efficiency of large numbers of individuals carrying out a variety of tasks. Such a criterion is not available, and the "construct validity" of a measure of effective intelligence must be examined by diverse approaches.

Relationship of test scores to age. Two studies with data relevant to the relationship between test scores and CA can be cited. In the first, Form B of the CSCT was administered to all the children in two representative schools in Davidson County, Tennessee, a metropolitan community, who were within three weeks of being 9–6, 10–6, and 11–6 years of age. Table 1 contains the data for the 66 children in this study. As may be seen, the two older groups of the study are rather small. If they are combined and compared with the younger group, the difference in mean scores is significant well beyond the .001 level.

A second study of the relationship of CSCT scores to age was carried out with a group of 50 orphanage children, again using Form B of the CSCT. These children ranged in CA months from 97 to 154, with a mean of 129.66 and a sigma of 15.7. The obtained correlation between chronological age and CSCT scores is significant at the .0001 level, but this is not very informative. The size of an *r* between chronological age and any measure of mental age will be greatly influenced by the range and distribution of chronological ages included. What is really needed, therefore, is an indication of the *r*s between CAs and mental ages on other well-known tests within this same group. This is available on this group of 50 children for the Goodenough Draw-A-Man Test, but for the Binet, only 33 of the 50 subjects were tested. All of the tests—Form L of the Binet, the Goodenough, and the CSCT—were administered within a period of five or six weeks. The Goodenough drawings were collected and scored by an advanced graduate student, and the same student administered the CSCT, which was, however, scored by the author. In Table 2 the correlations of each of the tests with ages in months are shown. In computing these correlations, mental ages were used for the Binet and Goodenough data.

Relationship of test scores to scholastic achievement. A widely used criterion of an intelligence test's validity is the correlation of the test with school achievement. Form B of the CSCT and Forms L and M of the Binet were given, plus the Stanford Achievement Test, to a group of 67 orphanage children aged 9–13, none of whom were in the study described above. As a group these children were retarded, as shown by a Binet IQ range (L and M were averaged) of 67.5 to 98. Table 3 summarizes the pertinent data on this group. The range of intellectual ability and of scholastic attainment within this group was rather narrow, considerably less than would be found within a single grade in the typical school system. In spite of this, both the Binet and the CSCT had substan-

Table 3

Range and Variability of an Orphanage Group on Two Tests of Intelligence and an Achievement Test

Test	Range	Mean	Sigma
Binet (L and M) MA	140–76	107.36	13.87
CSCT (Form B)	69–10	39.86	10.81
Stanford Achievement (grade equivalent)	5.4–1.5	3.22	.87

tial correlations with the Stanford scores. The Binet correlation was .623 and the CSCT correlation was .542. Since the Binet MAs and the CSCT were not highly correlated with each other (.349), a multiple *r* between these two and the Stanford Achievement was computed and was found to be .713. While the Binet was slightly more successful than the CSCT in predicting Stanford Achievement scores, it should be noted that the Binet scores involved were averages of two forms, requiring a minimum testing time of two hours for each child, while the CSCT required about fifteen minutes for each child.

Relationship of test scores to IQ gains in therapy. One other set of correlations may have particular interest for clinicians. As part of a larger experiment designed to increase IQs, 55 of these orphanage children with below-average IQs were exposed to an 18-month program of group play therapy, with 4 children in each group, and of enriched cultural opportunities. Both L and M of the Binet were given prior to and again at the conclusion of therapy. The range of IQ gains was not spectacular, from − 3.5 to + 17, with a mean of + 8.26 and an *SD* of 4.15. There was no relationship (.014) between the original IQ level and the amount of gain. However, the original MA level correlated .172 with the IQ gains and the CSCT scores correlated .295 with these gains. These *r*s are significant at the .11 and the .02 levels, respectively. If other corroboration of this finding is forthcoming, it is potentially important, for any brief procedure which can help to predict the gains of retarded children from therapy is worth knowing about.

Summary

A method of testing "effective intelligence" by an unstructured oral sentence-completion technique (Copple Sentence Completion Test) has been described. A median interscorer reliability of .97 was found in an investigation involving 10 scorers, and in other, smaller projects, the results were comparable. Split-half reliabilities for three studies reported in the present paper were .94, .95, and .93.

Three approaches to the test's validity were reported:

1. CSCT scores compared favorably with Binet and Goodenough mental ages in their correlation with chronological age.
2. CSCT scores compared favorably with Binet mental ages in predicting Stanford Achievement Test scores.
3. CSCT scores were significantly related to IQ gains subsequently made by a group of retarded children undergoing play therapy.

Received November 1, 1955.

References

1. Ebbinghaus, H. Ueber eine neue Methode zue Prufung Geistiger Fahigkeieten und ihre Anwendung bei Schulkindern. *Z. Psychol. Physiol. Sinnesorgane,* 1897, 13, 401–459.
2. Terman, L. M., & Merrill, Maud A. *Measuring intelligence.* Boston: Houghton Mifflin, 1937.
3. Trabue, M. R. Completion-test language scales. *Teach. Coll. Contr. Educ.,* 1916, No. 77.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

A Note on Porteus Maze and Wechsler-Bellevue Scores as Related to Antisocial Behavior[1]

Kenneth Purcell[2]

Mental Hygiene Consultation Service, Fort Dix, New Jersey

As part of a study on antisocial behavior and some of its personality correlates, data were collected on Porteus Maze and short-form Wechsler-Bellevue performance. Previous studies (1, 2, 6, 9), using the qualitative method for scoring the Maze Test described by Porteus (5), have consistently reflected differences between delinquents and nondelinquents. Nevertheless, it seemed worth while to include the Maze Test in the present investigation, mainly because the type of population used differed in important respects from earlier populations studied. Each of the former investigations had compared institutionalized delinquents with noninstitutionalized, presumably normal nondelinquent youths. The population employed in this study consisted of noninstitutionalized young male adults, all of whom were referred to a mental hygiene clinic for psychiatric problems. Porteus (7) has cited evidence to show that the Maze Test successfully predicts general social adjustment and adaptation. The question posed in this paper concerns the capacity of the Maze Test, qualitatively scored, to distinguish poor social adjustment characterized primarily by antisocial behavior from other types of poor social adjustment.

Hunt and French (3), in one of their reports on the CVS abbreviated intelligence scale, found a tendency for psychopaths and disciplinary cases to score lower on Comprehension than on Similarities and Vocabulary. However, the differences found were not statistically significant. It is suggested by the authors that this may well have been because of inadequate criteria for defining a psychopathic population. Since the groups used in this study were reliably separated along the dimension of antisocial behavior, a further examination of CVS results seemed appropriate. In addition, on the basis of our experience at this clinic, it was felt that the Block Design subtest generally offered the most valid estimate of intellectual capacity among antisocial subjects. This appeared to be because negativism and resistance were minimal and interest maximal for this subtest which Wechsler (8) reports as having one of the highest correlations with Full Scale IQ of all subtests (.73).

Method

A total of 56 male Army trainees between the ages of 17 and 26 with a mean age of 20 were used as *S*s for this experiment. All were psychiatric referrals to the mental hygiene clinic for whom psychological testing was requested. The only patients excluded were those suffering from organic brain damage or those diagnosed as psychotic.

It was the intent of this study to relate psychological test results to overt behavior. Therefore, each *S* was rated on manifest antisocial behavior trends on the basis of social history data routinely obtained by one of the psychiatric social work specialists. Both military and civilian data were used in making this evaluation. Originally two raters independently judged the degree of antisocial behavior on a seven-point scale with a reliability of .91 being obtained. This scale was later consolidated into a three-point scale to

---

[1] The author wishes to express his appreciation to Carl Eisdorfer, Marvin Greenstein, Maurice Kouguell, and Dr. Philip Kubzansky who collected data for this study and assisted in its statistical evaluation.

[2] Now at the University of Kentucky.

Table 1

Analysis of Variance of Qualitative Maze Scores of Three Groups Differing in Antisocial Behavior

Group	Mean*		Variance†	F	p
Group I	30.38	Between groups	3,543.97	3.39	.05
Group II	42.94	Within groups	1,044.15		
Group III	58.57				

* Ns = 24, 18, and 14 for Groups I, II, and III, respectively.
† Between groups $df = 2$; Within groups $df = 54$.

secure an adequate number of *S*s for each category. Group I, the non-antisocial group, was composed of *S*s whose overt aggressive behavior was minimal with no record of military or civilian disciplinary infractions. Group II, the intermediate group, consisted of *S*s who may have been in fights with moderate frequency and whose record was marked by some school truancy, perhaps minor infractions of the law, and AWOL from the military. Group III, the antisocial group, consisted of *S*s with an extensive pattern of delinquency including some criminal actions such as assault, rape, burglary, dope peddling, etc.

The three groups were compared on qualitative Maze Test scores and on the four Wechsler-Bellevue subtests noted above. In each case these measures were administered as part of a battery of clinical tests.

Results and Discussion

Analysis of variance of the qualitative maze scores, as shown in Table 1, discloses a difference between the mean *Q* scores (total weighted qualitative score) of the three groups significant at the .05 level. This is not so significant a result as has been reported in other studies of the Maze Test. For example, Docter and Winder (1) report a difference between delinquents and nondelinquents significant at the .0001 level. There are at least two reasons for the apparent insensitivity of the maze in differentiating between these two groups in the present study. First, the range of *S*s was marked by more continuity with respect to antisocial behavior than was true in other investigations. The presence of a group intermediate between the extreme antisocial and non-antisocial groups created additional overlap of scores. Second, the number of *S*s used in this study, particularly if one considers only the two extreme groups, was considerably smaller than the *N*s reported by other authors (1, 2, 5, 6, 9). If similar differences were obtained with larger *N*s, the significance level would, of course, be higher.

It will be remembered that Group I *S*s were also psychiatric referrals, the great majority of whom had manifested some form of social maladjustment in the Army situation. They may have been seclusive, overdependent, easily frightened by authority or socially deficient in almost any respect except that none of these *S*s were aggressively antisocial in their overt behavior. Since the *Q* scores of these *S*s were only slightly higher, and on a relative basis not at all higher than the *Q* scores of normal controls used in previous studies, it is apparent that the qualitative Maze Test tends to be sensitive only to certain types of social maladjustment. It seems to be maladjustment marked by nonconforming impulsivity that is singled out by this qualitative analysis. Impulsivity is probably the essential ingredient for high *Q* score. For example, one of the severely antisocial *S*s was a shrewd, calculating relatively controlled person whose record reflected many of these characteristics. He had been arrested on numerous occasions but always managed to elude conviction. His *Q* score of 5 was one of the lowest for all *S*s. In his behavioral notes the tester noted that *S* was calm, detached, and obviously out to demonstrate that he (*S*) could easily master the situation.

Since several authors (1, 7) have noted a small but significant correlation between *Q* score and intelligence, the intellectual factor must be measured and accounted for in some way. The task of estimating the intellectual capacity of antisocial *S*s is difficult since their test performance is so colored by uncooperative, negative attitudes. This was a particularly important problem in this study, as many of the *S*s were actively engaged in intense rebellion against any form of military authority.

Table 2 presents the results of the analysis of variance of the Wechsler-Bellevue data. The differences between the means of the three groups are significant for Prorated IQ, Similarities, Vocabulary, and Comprehension scores with antisocial *S*s achieving the lowest scores in each instance. By contrast, the differences found in Block Design means do not even approach significance. The reasons for this finding are at least twofold. First, almost every antisocial *S* had been an habitual truant and had left school much earlier than the non-antisocial *S*s. The benefits of educational experience on such tests as Vocabulary or Comprehension were lost to these *S*s. Second, the verbal interaction required by the verbal subtests reproduces former unpleasant school situations and creates many more opportunities for the arousal of authority conflicts than occurs during administration of the Block Design test. During the latter, *S* may focus his attention on impersonal objects without the necessity of replying directly to the examiner. Antisocial individuals seem to find it less necessary to assert their rebellious nonconformity in this more oblique relationship with the examiner.

Turning briefly to a comparison of Comprehension scores with those of Similarities and Vocabulary, we find, from an inspection of means and significance levels indicated in Table 2, the same tendency for antisocial *S*s to score relatively lower on Comprehension, as was noted by Hunt and French (3). Although the differences between these subtests do not attain statistical significance, the qualitative variation between the groups in their responses to certain items of the Comprehension test seems worthy of note. The answers of antisocial *S*s to the questions on the envelope, bad company, taxes, and law, commonly contained antisocial content in some part of the response, although giving credit to the response might be required under the rules of scoring. Perhaps a revised scoring scheme or the remodeling of Comprehension to encourage antisocial answers in

Table 2

Analysis of Variance of Wechsler-Bellevue Subtest Scores of Three Groups Differing in Antisocial Behavior

Variable	Mean*		Variance†	*F*	*p*
Prorated IQ					
Group I	103.45	Between groups	3,296.47	9.36	.001
Group II	96.26	Within groups	352.26		
Group III	76.35				
Block Design					
Group I	9.62	Between groups	8.14	.65	Non-signif.
Group II	8.94	Within groups	12.40		
Group III	8.28				
Comprehension					
Group I	10.22	Between groups	113.06	10.77	.001
Group II	8.39	Within groups	10.49		
Group III	5.00				
Vocabulary					
Group I	9.34	Between groups	50.32	5.22	.01
Group II	8.93	Within groups	9.64		
Group III	6.56				
Similarities					
Group I	11.00	Between groups	101.72	7.47	.01
Group II	9.81	Within groups	13.61		
Group III	6.83				

* *N*s = 24, 18, and 14 for Groups I, II, and III, respectively.
† Between groups $df = 2$; Within groups $df = 54$.

accordance with the suggestion of Hunt and French (3) would improve its discrimination along the dimension of antisocial trends.

Summary

One of the questions with which this study was concerned was whether the Porteus Maze test, qualitatively scored, could successfully distinguish between poor social adjustment marked by antisocial behavior and various other types of poor social adjustment. The second aspect of the study dealt with scatter scores among the Comprehension, Vocabulary, Similarities, and Block Design subtests of the Wechsler-Bellevue Intelligence Scale.

Fifty-six young adult Army trainees, all of them referrals to the Mental Hygiene Consultation Service, serve as *S*s. The *S*s were divided into three groups on the basis of wide variations in overt antisocial behavior.

Qualitative Maze Test scores differentiated significantly between three groups, the most antisocial group having the largest *Q* score. It appeared that impulsivity was the basic personality variable measured by qualitative evaluation.

Finally, our impression that Block Design performance would be the subtest least impaired by antisocial tendencies was confirmed. The results showed no significant differences between the three groups in Block Design scores, although the Comprehension, Vocabulary, and Similarities scores differed significantly. An explanation for these findings is offered in terms of the more remote examiner-subject relationship during Block Design administration as compared with the other three subtests.

Received December 20, 1955.

References

1. Docter, R. F., & Winder, C. L. Delinquent vs. nondelinquent performance on the Porteus Qualitative Maze Test. *J. consult. Psychol.*, 1954, **18**, 71–73.
2. Grajales, M. C. Porteus' Qualitative Maze Test as a measure of delinquency. Unpublished master's thesis, Fordham Univer., 1945.
3. Hunt, W. A., & French, Elizabeth G. The CVS abbreviated individual intelligence scale. *J. consult. Psychol.*, 1952, **16**, 181–186.
4. McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.
5. Porteus, S. D. *Qualitative performance in the maze test*. Vineland, N. J.: Smith Printing House, 1942.
6. Porteus, S. D. Q-scores, temperament and delinquency. *J. abnorm. soc. Psychol.*, 1945, **21**, 81–103.
7. Porteus, S. D. *The Porteus Maze Test and intelligence*. Palo Alto, Calif.: Pacific Books, 1950.
8. Wechsler, D. *The measurement of adult intelligence*. (3rd Ed.) Baltimore: Williams & Wilkins, 1944.
9. Wright, Clare. The qualitative performance of delinquent boys on the Porteus Maze Test. *J. consult. Psychol.*, 1944, **8**, 24–26.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

MMPI Differences Between Parents of Stuttering and Nonstuttering Children[1]

Leonard D. Goodstein
State University of Iowa

and W. Grant Dahlstrom
University of North Carolina

While there has been much published speculation and theorizing about the etiology of stuttering (e.g., 1, 8), until recently there have been relatively few empirical studies of the onset of this fairly common clinical problem. Darley (2), Johnson (9), LaFallotte (12), and Moncur (13, 14) have all attempted to investigate the circumstances surrounding the development of stuttering by intensively interviewing the parents of stuttering children. The present study represents a cross validation and extension of these prior investigations and includes an objective measure of parental adjustment, the Minnesota Multiphasic Personality Inventory (MMPI). This paper reports the MMPI results, while the more general findings are reported elsewhere (10).

The MMPI has been used in a similar way for evaluating parental adjustment in its relation to child behavior by Williams (18) in a study of families in which there was a cerebral-palsied child, and by Grossman (7) in relation to stuttering. The present study is an extension of this method to a larger, more carefully selected set of parents.

Subjects and Procedures

A family was considered for inclusion in the stuttering or experimental group if one or both of the parents regarded the child as a stutterer and if the duration of the stuttering was reported as less than 36 months. A family was considered in the nonstuttering or control group if the child had never been regarded as a stutterer by either of the parents. Each experimental family was matched to a control family on the basis of the child's age and sex as well as the socioeconomic status of the parents. Only essentially physically normal, white children were included in the study. The experimental subjects (*S*s) had been referred to the experimenters by speech therapists, physicians, etc., while the control *S*s were secured by appeals for volunteers to PTA's, church groups, etc. The complete details of subject selection are available elsewhere (10).

The routine administration of the booklet form of the MMPI was part of the standard interview procedure (see 10) that was followed with each parent who had agreed to participate in the over-all study. Usually one parent took the test while the other parent was being interviewed. If the tests were not completed before the interviewer was ready to leave, arrangements were made to have the booklets and answer sheets picked up at some later time.

While it had been planned to secure an MMPI profile from each parent, this goal was not realized due to a number of uncontrollable factors such as lack of time, refusal to cooperate, etc. MMPI profiles, however, were obtained from 224 of the 300 experimental parents and from 223 of the 300 control parents. An inspection of the other data did not suggest that those parents who had failed to complete the MMPI were different in any obvious ways from those parents for whom profiles were secured.

---

[1] This research was supported by a grant from the Louis W. and Maud Hill Family Foundation of St. Paul, Minnesota. The authors wish to thank Dr. Wendell Johnson, the principal recipient of the grant, for his help and encouragement during the execution of the research and the preparation of this manuscript.

All MMPI answer sheets were hand scored by the interviewers for the usual four validity and nine clinical scales as well as four special scales that were thought, on an a priori basis, to represent variables important to the onset of stuttering. These four special scales were the Taylor Manifest Anxiety Scale (*A*) (15), the Achievement Scale (*Ac*) (5), the Dominance Scale (*Do*) (6), and the Status Scale (*St*) (4). All profiles were also coded into Welsh's high-point coding system (16). For ease in tabulation and statistical analysis, the scores on the 17 scales as well as the high-point codes were then coded into a predetermined coding system and punched on standard IBM punch cards.

Results

Four groups of 100 profiles each, experimental mothers, experimental fathers, control mothers, and control fathers were selected from the 447 available MMPI profiles by excluding the 47 profiles whose validity scale scores suggested a significant degree of dissimulation. The cutting scores for exclusion were: *L*-scale raw score of 7 or higher, or *F*-scale raw score of 17 or higher, or *K*-scale raw score of 28 or greater. The numbers of cases excluded on the basis of these criteria were 17, 22, and 8, respectively; there were no obvious differences between the control and experimental parents with respect to these validity criteria.

The means and standard deviations on the 17 MMPI scales for the experimental and control parents, together with the significance levels of the differences between the experimental and control means are presented in Table 1. For ease in communication the scores in Table 1 are reported in values based on the original distributions of scores rather than the coded score distributions. The reported values are, consequently, approximate rather than exact; all statistical tests, however, were based upon the more exact coded score values.

A prior comparison of the means of the experimental mothers with the experimental fathers and of the control mothers with the control fathers gave no evidence of any sex differences in the scores and the data from the two sexes were combined. The only exception to this finding was for the *Mf* scale on which both groups of fathers had significantly higher ($p < .01$) mean T scores than both groups of mothers; there were, however, no significant differences between the experimental and control groups and the data were therefore combined in this case as well.

There were only three significant mean differences between the experimental and the control parents on the 17 MMPI scales. The control parents had significantly higher mean scores on the *K* and Dominance (*Do*) scales ($p < .05$, $p < .01$, respectively) while the experimental parents had a significantly higher mean score on the Anxiety (*A*) scale ($p <$

Table 1

Means and *SD*s* of the MMPI Scales for the Two Groups and the Significance Levels of the Differences Between the Means ($N = 200$ in each cell)

	Experimental parents		Control parents		
Scale†	Mean	*SD*	Mean	*SD*	p‡
?	6.5	7.9	5.8	4.6	ns
L	3.3	1.9	3.6	1.8	ns
F	3.2	2.5	2.8	1.8	ns
K	15.5	5.2	16.9	4.8	.05
Hs	51.6	8.7	53.0	7.4	ns
D	53.9	10.3	52.4	9.2	ns
Hy	55.5	7.9	56.1	7.7	ns
Pd	53.2	9.6	54.8	9.0	ns
Mf§	52.5	10.1	52.7	10.1	ns
Pa	52.8	7.8	52.3	6.8	ns
Pt	53.0	8.6	51.9	7.7	ns
Sc	51.7	9.0	52.2	7.8	ns
Ma	51.9	10.1	51.2	9.1	ns
A	14.3	7.6	11.8	6.9	.01
Ac	16.9	2.9	16.8	2.6	ns
Do	17.3	3.3	18.8	3.3	.01
St	21.4	3.9	22.5	4.0	ns

* These values are approximate as they were derived from the coded scores.

† The scores on the clinical scales (*Hs, D, Hy, Pd, Mf, Pa, Pt, Sc,* and *Ma*) are in T-score units; all the others are expressed in raw score units. *K*-scale corrections included where appropriate.

‡ The t tests for mean differences as well as the significance levels for these differences were based upon the more precise coded score values rather than the values reported in this table.

§ Both groups of fathers were significantly higher ($p < .01$) than both groups of mothers; there were, however, no significant differences between the stutterers and nonstutterers.

Table 2

Relative Frequencies of the Initial Scale in the High-Point Code for Each Group

Group	MMPI Clinical Scales*									Total
	Hs	*D*	*Hy*	*Pd*	*Mf*	*Pa*	*Pt*	*Sc*	*Ma*	
Experimental mothers	4	13	18	12	15	12	8	6	12	100
Experimental fathers	5	17	16	8	23	4	5	4	18	100
Subtotal	9	30	34	20	38	16	13	10	30	200
Control mothers	3	18	27	22	6	7	7	5	5	100
Control fathers	1	10	17	18	27	4	5	3	15	100
Subtotal	4	28	44	40	33	11	12	8	20	200
Grand total	13	58	78	60	71	27	25	18	50	400

* Only the nine clinical scales were used in this analysis.

.01). None of the other mean differences were significant at an acceptable level of confidence.

The relative frequencies in the four groups of *Ss* of the initial scale in the high-point code, that is, the clinical scale with the highest *T*-score value in the individual profiles, are presented in Table 2. The differences between the obtained frequencies were then analyzed by means of the chi-square test for a two-by-nine table. There were no significant differences between the experimental and control fathers, between the experimental and control mothers, or between the experimental and control parents combined over sex. There were, however, significant differences between the total male group and the total female group when the experimental and control groups are combined ($p < .001$). It may be seen in Table 2 that the *Mf* and *Ma* scales occurred relatively more frequently in the initial position in the male codes while the *Pa* and *Sc* scales were the more frequent initial points in the female codes.

The relative frequencies in the four groups of *Ss* of the secondary scale in the high-point code, that is, the clinical scale with the second highest *T*-score value in the individual profiles, are presented in Table 3. The differences between the obtained frequencies were again analyzed by the means of the chi-square test for a two-by-nine table. Again there were no differences between the experimental and control fathers, the experimental and control mothers, or between the total experimental and the total control group. Again there were significant differences between the total male group and the total female group when the experimental and control groups are com-

Table 3

Relative Frequencies of the Secondary Scale in the High-Point Code for Each Group

Group	MMPI Clinical Scales*									Total
	Hs	*D*	*Hy*	*Pd*	*Mf*	*Pa*	*Pt*	*Sc*	*Ma*	
Experimental mothers	9	11	21	12	6	14	12	9	6	100
Experimental fathers	4	18	13	14	15	11	8	5	12	100
Subtotal	13	29	34	26	21	25	20	14	18	200
Control mothers	9	6	20	14	13	12	8	7	11	100
Control fathers	2	13	18	16	16	9	6	9	11	100
Subtotal	11	19	38	30	29	21	14	16	22	200
Grand total	24	48	72	56	50	46	34	30	40	400

* Only the nine clinical scales were used in this analysis.

bined ($p < .02$). It may be seen in Table 3 that the *D*, *Mf*, and *Ma* scales occurred relatively more frequently in the secondary position in the male codes while the *Hs* and *Hy* scales were the more frequent secondary scales in the female codes.

The number of parents in the two groups with one or more *T* scores above 70 on the clinical scales (the usual cutting score) was then obtained in an effort to evaluate the relative frequency of "severe disturbance" in the two groups. There were 47 persons with one or more *T* scores above 70 in the control group and 49 persons in the experimental group; by eliminating from consideration those cases where both parents in the family had scores above 70, these numbers were reduced to 41 and 42, respectively. An inspection of the high-point codes of those "severely disturbed" individuals did not reveal any differences between the experimental and control groups although the sex differences that were discussed above (see Tables 2 and 3) were again present.

In an attempt to evaluate the differences in the severity of stuttering as a function of the "abnormality" of the MMPI scores, the experimental group was then divided into two subgroups, those families in which one or both parents had one or more clinical scale *T* scores above 70 (the disturbed group) and those families in which neither parent had a clinical scale *T* score above 70 (the nondisturbed group). Independent ratings by the interviewers of the experimental children on the Iowa Scale for Rating the Severity of Stuttering (11, pp. 129–136) were available for 38 children whose parents were placed in the disturbed group and 77 children whose parents were in the nondisturbed group. The mean rating for the children of the disturbed parents was 1.97 ($SD = .97$) while the mean rating for the children of the nondisturbed parents was 3.13 ($SD = .71$); the differences between these means is highly significant ($t = 6.4$, $p < .001$) and indicates that the children of the nondisturbed parents were the *more* severe stutterers. (A rating of 1 on the Iowa Scale indicates normal, fluent speech while a rating of 3 indicates a relatively mild degree of stuttering.)

Discussion

The most striking finding of this portion of the total study was the failure to establish any major differences between the parents of the stuttering and nonstuttering children. Although certain differences between the two groups of parents were statistically significant and psychologically provocative, the over-all impression left by the results is that whatever important differences exist between these groups is not reflected in their MMPI scores. Certainly the small absolute differences between the means and the large amount of overlap between the two distributions of scores would prohibit the use of the MMPI as an instrument for differentiating these groups in clinical work.

These results are in essential agreement with those of Grossman (7) who reported that the parents of stutterers gave more atypical or pathological responses on the MMPI than the parents of nonstutterers on the F scale but did not differ on any of the other scales. Her results were obtained from older, more chronic stutterers and her parent groups involved only an *N* of 42 each. It should be noted that the differences between her two groups were small in absolute magnitude and statistically unreliable.

It is essential to note that not only are both sets of parents very similar to each other in their mean MMPI profiles but they are also very similar to the original standardization group of Minnesota normals (17). The highest clinical scale mean for either group is 56.1 for the control parents on the *Hy* scale; this represents a deviation of only 0.6 *SD*s from the expected score of a group of psychiatrically normal persons, that is, 50.0. The highest clinical scale mean for the experimental group, who presumably should be more maladjusted, is 55.5, also on the *Hy* scale. These, of course, are the largest differences from the expected normal mean score and all the other obtained means are closer to the expected value of 50.0. While some of these differences may be of statistical significance, they are not, in all probability, psychologically important. It is more reasonable to ascribe such small differences to sampling factors than to any "real" differences between the individuals studied in the present investigation and the standardization group. These differences certainly are not evidence for regarding either the control or experimental groups of parents as psychiatrically abnormal or more maladjusted than adult persons in general.

The absence of abnormality in these mean profiles is even more strikingly brought home when these profiles are compared with the mean profiles of various neurotic and psychotic groups (17) and of college students (3, 17). Each of these latter groups have characteristic mean profiles with very marked peaks and valleys differing greatly from the flat mean profiles obtained from these parents. Such comparisons strongly support the contention that these individuals are, for the most part, essentially normal persons who differ little, if at all, from the group of normal adult Minnesotans who comprised the standardization sample. These results incidentally, especially those with the control group, offer strong support for the continued use of the MMPI as a screening instrument with presumably normal adults.

While the present results offer no evidence for concluding that these two groups of parents are very different from each other or from normal individuals generally, it should be recalled that 3 of the 17 mean differences, although the absolute differences were quite small, were significant and deserve further comment. On the *K* scale the control parents' mean was significantly higher than that of the experimental parents. This would indicate that the control parents were somewhat more defensive, less likely to admit psychological defects or problems, than the experimental parents. These control parents had not been asked, as had the experimental parents, to admit *as part of the selection procedure* that they regarded their children as deviant, and certainly might be expected to be less open and more defensive in such a clinical situation.

The finding that the control group mean on the Dominance (*Do*) scale was significantly higher than that of the experimental group, suggesting greater dominance in face-to-face personal relationships, is rather difficult to interpret and fit into the other findings and conclusions about the onset of stuttering. Little is known about the interpretation and usefulness of the *Do* scale as it is a recent addition to the MMPI scales and perhaps any more elaborate interpretation of this result should be postponed until additional information about the *Do* scale is available.

The experimental parents were significantly higher on the Anxiety (*A*) scale than were the control parents. Certainly the admission of greater tension and anxiety on the part of the stutterers' parents may be regarded as an important finding, but great care should be exercised in relating this finding to the onset of stuttering. The heightened anxiety on the part of the experimental parents may be an important etiological factor in the development of stuttering, or may be a subsequent reaction by these experimental parents to the development of stuttering by their children. The present results do not permit a resolution of these alternative interpretations.

Perhaps the best evidence that the experimental group of parents was essentially normal from a psychiatric point of view was the finding that the "deviant" group of experimental parents, those with one or more *T* scores above 70, had children with little or no stuttering as compared with the nondeviant experimental group. On the other hand, the MMPI profiles of the parents of the more severe stutterers were almost uniformly within normal limits and gave no indication of severe psychological pathology of any sort. It may be that these deviant parents volunteered for participation not so much because they were concerned with nonfluencies in their childrens' speech, but rather with their own problems; they may have regarded participation in the experiment as one way of receiving help with these more general problems, rather than an opportunity to discuss their childrens' speech. Such an interpretation would indicate a need for caution on the part of the speech therapist when he does find a stutterer's parent with a deviant MMPI profile.

The MMPI results are in general agreement with those of the over-all investigation (10) in that they do not support the notion that the parents of stutterers are very different in their behavior and attitudes from the parents of nonstutterers. Rather these results support the suggestion that those group differences that are important in the etiology of stuttering are rather limited and probably quite specific to the speech situation and the stuttering phenomenon.

Summary

The Minnesota Multiphasic Personality Inventory (MMPI) was administered to 447 parents participating in an interview study of the onset of stuttering in an attempt to compare the relative psychological adjustment and personality characteristics of parents of stuttering children (experimental group) and parents of nonstuttering children (control group). Four groups of 100 MMPI records, experimental fathers, experimental mothers, control fathers, control mothers, were selected by excluding those records whose validity scores suggested dissimulation.

There were only three significant mean differences between the experimental and control parents on the 17 MMPI scales used in this study; the control parents had significantly higher mean scores on the *K* and *Do* scales while the experimental parents had a significantly higher mean score on the *A* scale. A comparison of the high-point codes of the experimental and control groups revealed no significant differences between these two groups. There were no significant differences in the number of parents in the two groups with one or more clinical scale *T* scores above 70, the usual cutting score indicating some degree of psychological pathology. The children of the experimental parents with *T* scores above 70 were shown to have been rated as having significantly *less* severe stuttering than the remainder of the experimental group children.

These results were interpreted as indicating that the performances of these two groups were not significantly different, nor was the performance of either group significantly different from the expected performance of psychiatrically normal persons. There was no evidence that the etiology of stuttering is rooted in gross psychological pathology of the stutterer's parents but rather the results are to be interpreted as supporting the notion that such attitudes and adjustments of the parents as may be involved in the etiology of stuttering are probably limited and specific to the speech situation.

Received January 3, 1956.

References

1. Barbara, D. A. *Stuttering. A psychodynamic approach to its understanding and treatment.* New York: Julian Press, 1954.
2. Darley, F. L. The relationship of parental attitudes and adjustments to the development of stuttering. In W. Johnson (Ed.), *Stuttering in children and adults.* Minneapolis: Univer. of Minnesota Press, 1955.
3. Goodstein, L. D. Regional differences in MMPI responses among male college students. *J. consult. Psychol.*, 1954, **18**, 437–441.
4. Gough, H. G. A new dimension of status. I: Development of a personality scale. *Amer. sociol. Rev.*, 1948, **13**, 401–409.
5. Gough, H. G. The construction of a personality scale to predict academic achievement. *J. appl. Psychol.*, 1953, **37**, 361–366.
6. Gough, H. G., McClosky, H., & Meehl, P. E. A new personality scale for dominance. *J. abnorm. soc. Psychol.*, 1951, **46**, 360–366.
7. Grossman, Donna J. A study of the parents of stuttering and nonstuttering children using the MMPI and Minnesota Scale of Parents Opinion. Unpublished master's thesis, Univer. of Wisconsin, 1951. (Abstracted in *Speech Monogr.*, 1952, **19**, 193–194.)
8. Hahn, E. F. *Stuttering. Significant theories and therapies.* Stanford, Calif.: Stanford Univer. Press, 1943.
9. Johnson, W. A study of the onset and development of stuttering. In W. Johnson (Ed.), *Stuttering in children and adults.* Minneapolis: Univer. of Minnesota Press, 1955.
10. Johnson, W. (Ed.) *The onset of stuttering.* Minneapolis: Univer. of Minnesota Press, in press.
11. Johnson, W., Darley, F. L., & Spriestersbach, D. C. *Diagnostic manual in speech correction.* New York: Harper, 1952.
12. LaFallotte, A. C. A study of the parental environment of stuttering children. Unpublished doctor's dissertation, Univer. of Denver, 1948.
13. Moncur, J. P. Environmental factors differentiating stuttering children from nonstuttering children. *Speech Monogr.*, 1951, **18**, 312–325.
14. Moncur, J. P. Parental domination in stuttering. *J. Speech Hearing Disorders,* 1952, **17**, 155–165.
15. Taylor, Janet A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.
16. Welsh, G. S. An extension of Hathaway's MMPI profile coding system. *J. consult. Psychol.*, 1948, **12**, 343–344.
17. Welsh, G. S., & Dahlstrom, W. G. *Basic MMPI readings in psychology and medicine.* Minneapolis: Univer. of Minnesota Press, in press.
18. Williams, Janet T. A study of the parents of cerebral palsied and noncerebral palsied children using the MMPI. Unpublished master's thesis, Univer. of Wisconsin, 1951.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Response Bias and the MMPI

Eugene H. Barnes
VA Hospital, Downey, Illinois

In a recent article Berg (2) has proposed what he terms "the deviation hypothesis." This hypothesis is stated as follows: "Deviant response patterns tend to be general; hence those deviant behavior patterns which are significant for abnormality and thus regarded as symptoms, are associated with other deviant response patterns which are in noncritical areas of behavior and which are not regarded as symptoms of personality aberration" (2, p. 62). Deviant responses are defined as those which differ from modal responses or criterion group responses at a given level of significance. Berg states, "To test the hypothesis stated above, it is necessary only to identify the common or modal responses to any suitable series of stimulus patterns and predict that those subjects whose responses rather consistently go against the modal preferences will be deviant in the sense of exhibiting symptom patterns associated with abnormal states. It should be noted that content of stimulus patterns is not regarded as important" (2, pp. 62–63).

In a study pertinent to this hypothesis Barnes (1) used the Perceptual Reaction Test (PRT) as an instrument for eliciting deviant response patterns. The PRT is a series of sixty abstract designs made with ruler and compass to which the *S* is required to respond by marking one of the following options: *Like much, like slightly, dislike slightly,* or *dislike much*. By the use of deviant response patterns alone it was possible to construct scales for certain categories of psychopathology. In this study it was found that psychotic groups differed in their option preferences from normal groups in that the former generally preferred *like much* while the normal groups tended to prefer *dislike slightly*. A set to choose *dislike much* tended to be characteristic of character disorders in contrast to the performance of psychotic and normal groups. Cottle and Powell (4) demonstrated that answering all of the MMPI items "true" produced a profile with high scores on the psychotic triad while scoring all items "false" produced high scores on the neurotic triad. Berg cites these studies as evidence suggesting that, "a marked tendency to say 'yes' or 'agree,' etc., in personality inventories should be associated with psychotic states while 'no' or 'disagree' option choices should be associated with neurotic states" (2, p. 66).

There are evidently two parts to the deviation hypothesis as stated by Berg. The first and major part concerns merely the association of deviant response patterns on psychological tests with deviant behavior patterns. It asserts a general relationship between deviant test behavior and the deviant behavior of psychopathological conditions. This hypothesis draws together a variety of behavioral phenomena and makes explicit an assumption which has been implicit in much prior work. Few would argue that a "blood response" on the Rorschach was deviant behavior in a socially critical area, while few would disagree that strangling one's mother is. Nevertheless, a tenuous relationship has been assumed between "acting out" behavior and the socially nonsignificant response of "blood." The second and minor portion of the hypothesis concerns the directionality of such deviant response patterns as have been observed. This portion of the hypothesis asserts, in effect, that different types of response sets or biases are associated with different types of deviant behavior.

Individual record forms of the MMPI are a convenient source of material for testing certain deductions from the deviation hypothesis since the scoring system provides that only those items which are answered in the infrequent direction are recorded (6, p. 8). These items are, by Berg's definition, deviant responses. Thus, if content of stimulus patterns is not important, it would be inferred that the number of deviant responses, i.e., *X*s, on the record form of the MMPI should be as effective a measure of psychopathology in general as the profile itself. A scale of the MMPI which specifically purports to measure a psychotic reaction is the *Sc* scale. This scale, then, should correlate significantly with the total number of deviant responses. A preliminary check with grouped data revealed that the number of *X*s correlated .92 with the *Sc* scale for a sample of 24 male veterans in a neuropsychiatric hospital. The correlation to be expected from the number of common elements is .38 as calculated from a formula given by McNemar (8, p. 118). This limited confirmation suggested a more systematic exploration.

A large number of studies have indicated that the MMPI and the groups of scales known as the neurotic triad (*Hs, D, Hy*) and psychotic triad (*Pa, Pt, Sc*) have useful clinical validity (3). By inference from the deviation hypothesis, the total number of deviant responses should generally correlate with scales indicative of "symptom patterns associated with abnormal states" (2, p. 63). From the second part of the hypothesis it follows that deviant responses on the MMPI resulting from responding "true" when the modal response is "false" should correlate most highly with the scales of the psychotic triad while responding "false" when the modal response is "true" should correlate more highly with the scales of the neurotic triad. Since these tendencies to prefer certain option choices are conceived as being different response sets, the tendency to answer atypically true should not be correlated with the tendency to answer atypically false.

Procedure

Subjects. The *S*s were 40 males who, for one reason or another, had been given the MMPI by the author. These *S*s were a very heterogeneous group including psychotic outpatients, neurotics, vocational counseling cases, and court referrals. Variances on the MMPI clinical scales are, therefore, doubtless larger than from a random sample of the general population. It should also be noted that they were not representative of any homogeneous clinical grouping. The group was probably representative of those who come to the attention of a psychologist in private practice.

Method. Pearson *r*s were machine calculated for correlations of each of the clinical scales and three validating scales with, first, total number of deviant responses, second, number of deviant responses resulting from answering items false, and third, number of deviant responses resulting from answering items true.

Results

Results are summarized in Table 1. It will be noted that the differences in relative sizes of correlation coefficients on the neurotic triad and psychotic triad are all in the predicted direction, a result significant, without respect to magnitude, at less than a 2 per cent level of confidence as calculated by the sign test. The correlation between the deviant responses obtained by answering false and those obtained by answering true is less than the standard error and we may accept the hy-

Table 1

Correlations of MMPI Scales with Deviant Responses

Scale	Total number of deviant responses	Deviant responses answered true	Deviant responses answered false
L	−.09	−.24	.37
F	.74	.66	.43
K	−.60	−.69	.06
Hs	.66	.52	.56
D	.56	.34	.73
Hy	.12	−.09	.59
Pd	.71	.68	.28
Mf	.32	.28	.21
Pa	.60	.52	.37
Pt	.87	.84	.31
Sc	.93	.90	.36
Ma	.38	.50	.21
Dev. responses answered false		.11	

pothesis of no correlation. Certainly the total number of deviant responses is significantly associated with the clinical scales.

Discussion

This study was designed to test certain deductions derived from the deviation hypothesis and collateral research, not as a definitive test of the total hypothesis. Within this restricted frame of reference the deductions have been confirmed and this confirmation indirectly lends support to the hypothesis from which they were derived.

The deviation hypothesis does not take into consideration the matter of item content. If a preference for a response option exists in a normal population, deviating from this preference should add some increment to an indicator of psychopathology. In the present study we find evidence of such a phenomenon. With no attention given to particular items or their content, a mere summing of the deviant responses bears a striking relationship to certain of the MMPI clinical scales, as shown in Table 1.

One interesting implication of this portion of the hypothesis is that it would provide a theoretical rationale for the behavior of the *F* scale of the MMPI. This scale was constructed from items which were very rarely answered in the scored direction by normal *Ss* (5). On inspection these items seem to have no psychological coherence or communality, yet it has been observed that the *F*-scale score is elevated in personality aberrations (7, 9). The *F* scale has even been dubbed somewhat informally and inelegantly "the screwball scale."[1] The deviation hypothesis would predict just such a result from a scale so constructed. It would, in fact, predict that even a random selection of items scored in this manner would give a similar result.

With respect to the direction of deviation, it is clear that deviating by atypically answering true is associated with MMPI scales indicative of a psychotic type of pathology while atypically answering false is associated with scales indicative of a neurotic reaction. It may be argued that this result is obtained simply because the scales of the neurotic triad are more heavily weighted with items scored when they are answered false and those of the psychotic scales more heavily weighted with items scored when answered true. Although it is a fact that a degree of such differential weighting exists, such an "explanation" is circular in that it fails to explain the origin of such a weighting. Why did it happen, in the empirical derivation of the scales, that more items were found to discriminate between certain clinical groups and normals when answered in just these directions? In view of the results obtained in this study and the other studies cited, an answer in terms of chance distribution of item content seems untenable. It seems increasingly probable that response sets play a much more important role in personality testing than has been suspected.

Summary

A study was undertaken to test some deductions concerning the MMPI derived from Berg's deviation hypothesis (2). According to this hypothesis *Ss* whose responses consistently go against modal preferences should display symptom patterns of abnormality. Further, this hypothesis would predict that atypically responding true should be associated with psychotic states while atypically answering false should be associated with neurotic conditions. These two tendencies should be relatively unrelated to one another.

Since the record form of the individual MMPI provides that only items answered in the infrequent direction are entered, it is possible to count the number of deviant responses, to separate these into those obtained by answering atypically false and atypically true, and to correlate these scores with the scales of the MMPI. These operations were performed and it was found, as predicted, that the total number of deviant responses correlated highly with the clinical scales, that the tendency to answer atypically true was more highly correlated with the scales of the psychotic triad while the tendency to answer atypically false was more highly correlated with the neurotic triad, and

---

[1] If memory serves correctly the author first heard this label used by Dr. E. S. Bordin.

hat the two tendencies were not significantly correlated.

It is concluded that the deviation hypothesis has received some indirect confirmation both with respect to its general statement and with respect to direction of deviation. Also, it is concluded that response sets of this type play a larger part in personality testing than has been suspected.

Received December 1, 1955.

References

1. Barnes, E. H. The relationship of biased test responses to psychopathology. *J. abnorm. soc. Psychol.,* 1955, **51**, 286–290.
2. Berg, I. A. Response bias and personality: The deviation hypothesis. *J. Psychol.,* 1955, **40**, 61–72.
3. Cottle, W. C. The MMPI: A review. *Kan. Stud. Educ.,* 1953, **3**, No. 2.
4. Cottle, W. C., & Powell, J. O. The effect of random answers to the MMPI. *Educ. psychol. Measmt,* 1951, **11**, 224–227.
5. Hathaway, S. R., & McKinley, J. C. *Manual for the Minnesota Multiphasic Personality Inventory.* (Rev. Ed.) New York: Psychological Corp., 1943.
6. Hathaway, S. R., & McKinley, J. C. *Minnesota Multiphasic Personality Inventory:* Manual. (Rev. Ed.) New York: Psychological Corp., 1951.
7. Kazan, A. T., & Sheinberg, I. M. Clinical note on the significance of the validity score (*F*) in the MMPI. *Amer. J. Psychiat.,* 1945, **102**, 181–183.
8. McNemar, Q. *Psychological statistics.* New York: Wiley, 1949.
9. Schmidt, H. O. Test profiles as a diagnostic aid: The MMPI. *J. appl. Psychol.,* 1945, **29**, 115–131.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Delaying Capacity, Fantasy, and Planning Ability: A Factorial Study of Some Basic Ego Functions

Jerome L. Singer, Harold Wilensky, and Vivian G. McCraven
Franklin D. Roosevelt VA Hospital, Montrose, N. Y.[1]

An increasing ability to delay gratification and to inhibit a direct motor approach to a desired object appears to be a necessary feature in the development of an identifiable self in human personality. Intimately associated with this delaying capacity is the ability to resort to thought or ideation as "experimental action" (8) in a behavioral medium freed of temporal and spatial boundaries. As early as 1900 Freud (7) had suggested a theoretical linkage of delay and thought development and this relationship became a pivotal feature of his concept of the reality principle and, later the ego. Similar theoretical and empirical associations of thought and motor inhibition or delay have also been described by psychologists and psychiatrists in diverse fields (13, 19, 21, 30). With the increased attention paid by psychoanalytic theorists to ego psychology and to functions such as planning, synthesizing, and controlling motility (9, 11, 19), the possibility of integrating the psychoanalytic hypotheses with the concepts of psychologists involving the theoretical link between fantasy, language, and motor behavior emerges more clearly (22). The present study represents a further step in a program of research aimed at investigating some relationships of thought and delay in ego structure.

The specific approach to this problem has been through the medium of an intensive study of the Human Movement response (*M*) on the Rorschach inkblots and its behavioral correlates. Rorschach (20) pointed out that those persons who saw numerous human figures in motion on the blots tended to be inhibited in their own motor activity and given to considerable imagination or fantasy, "movements in the mind's eye." In effect, therefore, Rorschach's observations, and those of subsequent Rorschach workers (16, 18, 21) suggest that the *M* response may serve as a crude measure of delaying capacity and of imaginative tendencies which are basic ego-functions in Freud's formulation. The experimental work on the vicarious functioning of sensory and motor patterns carried out by Werner and Wapner and their colleagues is consistent with this approach and has carried it a step toward more precise formulation in perceptual theory (30, 31, 32). A number of studies in the present series have also supported Rorschach's observations by demonstrating that individuals whose Rorschachs contain relatively more *M* responses are capable of greater motor inhibition in slow-writing tasks, move about less in a waiting-room situation, and tend to tell more imaginative stories to TAT pictures (14, 23, 24). In the study to be presented here this work will be extended to include a large number of tests and behavioral situations which presumably tap aspects of impulse control or delaying capacity, planfulness, and fantasy. The general problem raised is whether consistent individual patterns emerge along a dimension that links delaying capacity, fantasy, planning ability, and various measures of motility control.

In approaching a concrete formulation of the general hypothesis a series of tests and behavioral indices were selected which, while superficially dissimilar, were expected to define a common fantasy-motility factor. The tests included a general intelligence measure, employed as a control since it was felt necessary to determine the extent to which the dimension in question was related to intelligence. The Rorschach test was included, with Rorschach's *M* response in particular expected to yield a high loading on a fantasy-motility factor. In order to evaluate the generality of the Human Movement phenomenon and to control for total number of Rorschach responses, Barron's Movement Threshold Inkblots (1) were included. Another estimate of fantasy tendencies was obtained from stories told by the *S*s to four Thematic Apperception Test pictures scored for "transcendence," i.e., introduction of characters, emotions, and temporal references which went beyond the mere description of the picture. Still another manifestation of fantasy tendencies characteristic of ego functioning might be an inclination toward setting high goals. A measure of level of aspiration was therefore included as well as a measure of sheer speed of performance in a simple motor task. The latter measure, however, would not be expected to show a significant loading on the factor in question. Another aspect of the adaptive nature of a developed capacity for fantasy was considered to be the ability to admit to awareness a painful experience. Some qualitative results in a previous study (33) had suggested interesting individual differences in admission of distress by *S*s following a failure experience which had measurably interfered with intellectual performance. A crude measure of admission or denial of disturbance following failure was therefore included in the test battery with the hypothesis that it would carry an appreciable loading on a fantasy factor. A somewhat similar rationale led to the inclusion of a technique which was felt to measure compliance with the examiner's wishes and persistence in a routine task to please another person despite psychological satiation. Finally, some attempt was made to obtain a measure of planning ability or foresight since this seems so clearly a characteristic of a well-developed fantasy-life. One approach involved self-ratings on a simple scale devised by Small (26). The other called for use of the Porteus Mazes, long described as measures of foresight by their originator (17), a claim afforded at least some support in a recent study by Small (26). The Porteus Mazes would appear particularly relevant since success in their execution involves the capacity both to delay motor reactivity, curbing impulsive solutions, as well as the more ideational component of visually moving through the mazes in advance of the pencil.

---

[1] The authors are indebted to Drs. John F. Winne and Jacob Cohen who gave valuable assistance in the statistical analysis of the data and to the psychology staff, trainees, and ward personnel who participated in some of the evaluative procedures.

Turning now to tasks and situations involving delaying capacity and control of motility, one possibility that suggested itself was a time-estimation procedure. Since the large majority of *S*s tend to respond too soon when instructed to indicate when a stated time interval has elapsed, sheer accuracy gave an adequate measure of delaying ability and control of impusive reaction. A better measure might have been to consider excessive delaying, but this occurred too rarely in prior trials of the procedure. A measure of motor inhibition, a slow-writing task originally devised by Downey, was included. A number of previous studies had suggested a consistently positive relationship between the ability to delay writing tempo and Rorschach's *M* response (14, 23, 24). To obtain some evaluation of longitudinal gross motor tendencies, ratings on certain variables were obtained from ward personnel who knew the *S*s for at least three months. Four variables based on the ward behavior items drawn from the Multidimensional Rating Scales (12) were employed. One involved aggressive behavior, frequently acting out a hostile impulse either through physical assaults or verbal imprecations. Another, cooperativeness with ward routine, presumably an example of controlled motility, would be expected to show a positive loading on a fantasy-motility factor. A measure of external interests, as manifested by behavior demonstrating some measure of orientation toward the hospital activities program and other people, was included. Finally, a series of items dealing with diffuse energy expenditure, gross motor activity not evidently hostile in intent, was employed, which might be expected to yield a negative or zero loading on a fantasy-motility factor.

In summary, then, the specific form of the investigation involved the prediction that a battery of procedures which on the surface measured a wide variety of performances ranging from fantasy capacity as seen in storytelling, perception of humans in activity on inkblots, planning in a maze test, slow-writing, and controlled motility, would all be revealed as aspects of an underlying, more general dimension in which the capacity for imaginal living and control of impulsive motility were linked. Since clinical experience and some previous research had suggested the value of Rorschach's Human Movement response as a measure of both the fantasy and motility aspects of the assumed underlying factor, it was anticipated that Rorschach's *M* response and the threshold for Human Movement on Barron's inkblots would show especially high loadings on this fantasy-motility factor.

Method

Subjects. There were 100 *S*s, all male veteran patients bearing the diagnosis of schizophrenia. They were chosen chiefly on the basis of their having recently received a Rorschach examination and an intelligence test. Their average age was approximately 31, mean date of first hospitalization 1949, mean IQ 108, and their years of education averaged 10.5. Social class background for this group was predominantly lower-middle and upper-lower and their ethnic backgrounds approximated the general hospital picture. Although all the patients in this study were unquestionably psychotic with hallucinations or delusional trends, only those *S*s who proved to be in sufficient contact to follow instructions for all of the tests were employed in the sample. It should be noted that the patients in this study represented a different sample of *S*s from those used in two prior investigations in this series (23, 24).

Procedure. A battery of special procedures was organized with an instruction booklet, and a scoring sheet.[2] With the exception of the Rorschach and Wechsler-Bellevue, which had usually been given some weeks or occasionally months earlier, all the data were obtained at one 90-minute sitting. The behavior ratings were obtained shortly after testing by interviews with ward personnel. Much of the testing was carried out by two of the authors (J. L. S. and H. W.) with occasional administrations of the battery by clinical psychologist trainees who had received instruction in the testing procedures.

Wechsler-Bellevue Intelligence Scale (Variable 1). For various practical reasons it was not possible to obtain full-scale IQs on all *S*s, but a prorated IQ estimate was obtained. These estimates were based on verbal tests usually, and only occasionally were performance test IQs available.

Rorschach. The Rorschachs were standardly administered examinations scored in accordance with Klopfer and Kelley (**10**). All protocols were rechecked by conference to eliminate the few questionable scorings of the original examiners. The remaining test battery was administered by the investigators without prior knowledge of Rorschach data for the *S*s in order to eliminate unconscious influence by the examiner. All determinants were scored but only those with sufficient frequency of occurrence and range in the sample were employed in the analysis. The Rorschach scores with Variable Numbers were: Total Responses (*R*, var. 2), Human Movement (*M*, var. 3), Animal Movement (*FM*, var. 4), Form-level (Beck's $F + \%$, var. 5), Black or Gray shading (Klopfer's *C′*, Beck's *Y*, var. 6), Form-Color (*FC*, var. 7), and Color-Form (*CF*, var. 8).

Barron's Movement Threshold Inkblots (Variable 9). This test consists of thirty black blots presented one at a time for thirty seconds with usual Rorschach instructions. Only one response per card is obtained and the score is based on the card in which the first clear *M* response is elicited. For purposes of simplifying the factor analysis, the scores on this variable were arbitrarily reflected prior to insertion in the matrix; a positive correlation in the matrix thus means that a low Movement Treshold is associated with a high score on the other variable.

Thematic Apperception Test Transcendence Index (Variable 10). Four cards of the TAT were administered in accordance with instructions developed in an earlier study (24). The cards were 1, 6BM, 7BM, and 12BM. The Transcendence Index, adapted from work by Weisskopf (29) called for scoring of items that involved addition of material not directly on the card, e.g., new characters, names, past or future events, etc.

---

[2] Copies of the correlation matrix are available on request from the authors. Tables giving the intercorrelation matrix, the transformation matrix, instructions for the various test techniques, and sample data sheet have been deposited with the American Documentation Institute. Order Document No. 4968 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Planning Self-Rating (*Variable 11*). This simple six-point rating scale was used by Small (26) in a study of planning tests. It calls for *S* to rate himself along a scale of planfulness in which each point is defined. The points range from: "I never plan things in advance," through "I always plan things out carefully in advance."

Porteus Mazes (*Variable 12*). This test was given in accordance with instructions in Porteus' Manual (17). Only the quantitative score (TQ) was employed here although further analysis by qualitative means is being planned.

Wechsler's Number-Square Test (*Level of Aspiration*). This procedure, modified from a technique devised by David Wechsler, called for *S* to draw squares around numbers printed serially in rows of ten across a page. Instructions called for speed and 30 seconds were allotted for a trial of 40 numbers. Prior to the first trial *S*s were asked to estimate their expected performance given only the information that "people do between 18 and 30 in half-a-minute." Following performance a second estimate was obtained. This procedure was carried out twice so that there were in all three expectation scores (one prior to performance) and two actual performance scores. Measures included Initial Expectation score (E_1, Var. 13) and Level of Aspiration $E_2 + E_3 - P_1 + P_2$ (L.A., Var. 15). The score obtained on initial performance (P_1, Var. 14) was also employed as a measure of sheer speed of reaction.

Time Estimation (*Variable 16*). This task called for *S* to indicate by saying "Stop" when he thought that a given time period had elapsed. Three times were used, 15, 30, and 60 seconds. Deviations from actual time were added to give a total deviation score. The measure used was *Accuracy*, based on absolute deviation.

Block Authority-Reaction Test (*Variable 17*). This procedure modified from a technique devised by the Blocks (2), called for *S*s to put ten pegs in holes and remove them at their own pace, repeating this procedure until their psychological satiation point was reached. The *S*s were informed that they could stop at will, but when they did so *E* asked, "Wouldn't you like to do some more?" and then, "Are you sure you'd like to stop now?" If *S*s continued the task in this ambiguous situation, they were scored for positive response to authority.

Motor Inhibition Test (*Variable 18*). This technique, derived originally from the Downey Will-Temperament Scale, had been used in a number of studies in this series (14, 23, 24). It called for *S* to write the phrase, "New Jersey Chamber of Commerce," as slowly as possible without stopping motion of the pencil. The score was the time in seconds taken to complete the phrase. The difference between an initial trial at normal tempo and the experimental trial was the actual measure.

Digit Frustration, Admission of Disturbance (*Variable 19*). This task, employed originally in a study on frustration by Wilensky (33), involved straight-forward digit retention except that *S* was taken beyond his limit of success, and his errors on the digit series beyond his memory span pointed out. Following a standard number of failures, the digit spans were presented in decreasing length until a long period of sustained success occurred. Evidence of disturbance was based on the difference between initial level and postfrustration level of digit retention. The variable employed here was the admission or denial of disturbance following failure. This was a dichotomous classification; those patients whose responses were unintelligible or who actually refused to respond were not considered, and consequently the *N* for this variable was lower than for the others.

Behavior Ratings. To obtain some evaluation of long-term motility patterns, items from the Multidimensional Rating Scales were grouped under four headings: those pertaining to *Aggression* (Var. 20), verbal or physical assaultive and destructive behavior on the ward, *Cooperativeness* (Var. 21) with ward routine, *Interest Level* (Var. 22), as seen in voluntary participation in ward activities, and *Diffuse Energy Level* (Var. 23), excessive motor activity not specifically hostile in intent. Ward personnel, aides, nurses, physicians, and ward clinical psychologists, all of whom had daily contact with the patients in question as well as with the hospital's daily listing of unusual behavior on each ward, were interviewed and each patient rated on a three-point scale for each item in the four categories. The rater was not cognizant of test results at the time of his interviews and ratings.

Results and Statistical Analysis

Following inspection of the various distributions, the scores on each variable were dichotomized at the median. In almost all cases good splits were obtained, the exceptions being the Authority-Reaction, where only 30% of *S*s met the criterion of compliance and the Energy-Level variable, where the cut was hampered by an extremely restricted range of scores. Following dichotomization, phi coefficients were obtained and transformed to estimates of Pearson's *r* in accordance with the procedure suggested by Edwards (5). A comparative sampling of tetrachoric coefficients revealed no sizable differences in the coefficients obtained by the two procedures—most deviations being of the order of 0.01 to 0.03. Recognizing the limitations of *r*s obtained in this fashion, the authors are following Wittenborn's reasoning (34) in attempting a factor-analytic study of Rorschach data.

A cluster analysis had led to the elimination of a number of variables which showed

Table 1

Centroid Loadings, Communalities, and Common Factor Loadings

Variables	I	II	III	IV	h^2	A	B	C	D
1. IQ (W-B)	27	26	−06	−17	17	01	25	19	−12
Rorschach Variable									
2. *R*	72	27	21	18	66	07	35	53	44
3. *M*	82	−67	25	−02	1.18	81	10	−27	65
4. *FM*	52	24	32	32	53	−07	27	52	52
5. *F*+%	32	−11	11	−17	16	22	21	−08	14
6. *C′*	25	55	−19	10	41	−19	14	57	−14
7. *FC*	46	29	−16	33	43	07	−03	56	15
8. *CF*	22	18	−16	28	18	02	−11	37	06
Experimental Battery									
9. Mt	60	−47	14	33	71	56	−14	01	62
10. TAT	58	27	30	06	50	−03	44	42	38
11. PSR	38	28	08	−47	44	02	55	07	−13
12. Port	45	−12	−28	−24	36	48	03	−07	−13
13. E_1	38	16	38	−52	58	−01	71	−06	09
14. P_1	31	42	21	−41	48	−18	64	18	−07
15. LA	11	−30	13	−11	14	23	04	−27	15
16. TIM	35	12	−18	12	18	17	−03	28	02
17. AR	24	−16	39	30	33	04	07	11	57
18. MIT	74	−25	−23	−10	67	69	00	00	12
19. AD	45	−26	36	20	44	24	12	05	60
20. Agg	19	36	−16	17	22	−11	02	44	−06
21. Coop	19	−54	−26	−34	52	64	−16	−53	−15
22. Ex I	28	27	−50	−49	64	23	16	04	−59
23. En L	15	33	−29	07	22	−04	−03	35	−21

negligible correlations in the matrix, e.g., Age, Date of First Hospitalization, Education (which correlated chiefly with IQ). The resulting matrix of 23 variables was then subjected to a factor analysis, using Thurstone's centroid method (27). Communalities were estimated by means of minor centroids. The relatively small residuals remaining in the third residual matrix and the unknown reliability of the method of estimating *r* warranted extraction of only four factors. The axes were rotated to oblique simple structure following Thurstone's method of rotation by two-dimensional sections. The centroid factors, the calculated communalities and oblique factor loadings are presented in Table 1. Since the oblique factors show slight intercorrelations, a second-order factor analysis was carried out. The matrix of intercorrelations of the factors, the second-order centroid factor loadings, and the rotated factor loadings are presented in Table 2.

Consideration of Factors A and D suggests at least partial support for the hypothesis of the study since Rorschach *M* and Movement-Threshold show sizable loadings on the two factors in which measures of controlled motility and fantasy are also appreciably loaded. Factor A derives by far its highest loadings from *M*, Movement-Threshold, Porteus Mazes, Motor Inhibition, and Cooperation in Ward Behavior. It thus appears to reflect controlled motility primarily and it seems possible that it represents the kind of motor restraint that would lead to compliance and inhibition of impulses. Factor D, on the other hand, with its sizable loadings on *M*, Movement Threshold, Admission of

Table 2

Intercorrelations Among Factors, Second-Order Centroid Loadings, and Oblique Rotated Loadings

Factor	A	B	C	D	I	II	h^2	AA	BB
A		40	44	−15	68	−16	49	15	45
B			12	−17	51	−39	41	−14	58
C				−22	58	29	42	50	00
D					−36	−19	17	−32	−01

Distress after Failure, Animal Movement, External Interests (negative loading), Continuation of Task in Response to Authority, Rorschach Responses, and Transcendence on TAT stories, appears to represent an ideational or introspective dimension of personality, particularly since all of the tasks involve little or no gross motor activity. Factor B might well be termed an ambition or need-achievement factor since it combines high loadings on productivity measures (Performance Speed, Rorschach Responses, and TAT Transcendence) with self-estimates of achievement potential (Initial Expectation on Number Square Test) and planfulness (Planning Self-Rating). Factor C with its high loading on Rorschach Color and Shading, Animal Movement, Rorschach Responses, Cooperation (Negative), Aggressive Ward Behavior, TAT Transcendence, and Energy Level in Ward Behavior, seems to represent a combination of productivity or drive with emotionality and uninhibited activity. The high loading of Rorschach *C′* and *FM* would appear to support certain views of Piotrowski (16), who has suggested that numerous *C′* and *FM* responses in Rorschach records are indicative of acting-out of impulses and animation, respectively. Factor C may well represent some aspect of a pressure for outward direction of living, perhaps a flight into activity and environmental manipulation which is manifested both motorically (high Aggression and Energy Level, low Cooperation) and ideationally (high TAT Transcendence and Rorschach Responses). It closely resembles a factor termed "Emotional Surgency" by Foster (6)[3] in a somewhat similar study of Rorschach data.

In the second-order analysis, Factor AA has a high positive loading for *Emotional Surgency* and a moderate negative loading on *Introspectiveness*. This result suggests at least some validity to Rorschach's original concept of the experience type. Factor BB is clearly defined by high loadings in *Controlled Motility* and *Ambitiousness* or *Need-Achievement*.

Discussion

The results of the present investigation, while grossly supporting the formulation advanced at the outset, make it clear that, at least for this sample, the initial hypothesis of a fantasy-motility dimension was overly simplified. The fact that Rorschach's Human Movement response showed significant correlations with almost all of the variables presumably measuring fantasy, planning or delaying ability, and motor inhibition is striking confirmation of the original formulation of Rorschach (20) concerning Human Movement and of subsequent interpretations of its theoretical meaning and behavioral implications (14, 18, 22, 30). Although Rorschach *M* and Barron's Movement-Threshold show high loadings on both Factors A and D, the emergence of two separate and relatively uncorrelated factors, one primarily involving motor control and the other ideational activity, suggests a more complex situation than was hypothesized. It seems possible that at least two types of persons are likely to present Human Movement responses to inkblots. One type consists of persons with considerable ability to restrain impulsive response and to plan ahead and cooperate for their own good in the numerous hospital situations involving waiting or complying with routine. The other group, while sensitive to the wishes of authority and perhaps willing to comply, is too withdrawn into introspection and self-evaluation or autistic reverie to make effective use of their potentialities for effective motor control and planfulness. This speculative distinction suggests more extensive qualitative exploration of the nature of Movement responses, perhaps to determine whether differences in content or elaboration of the percept may account for the association of *M* responses with the two separate patterns of response. Inasmuch as the two Human Movement measures have high loadings on both the Motor and Ideational factors in this study, the theory linking movement perception with controlled motility and fantasy receives some support. It may well be that the split factors which have emerged may be a consequence of the pathology of the patients and that study of a normal sample might clarify this issue considerably. Should fantasy-motility linkages emerge in other studies of this sort, the possibility of an operational formulation of the concept of ego functions may open important research avenues. At a theoretical level, it may become feasible to explore the nature of individual differences in thinking and motility as a function of family constellation or cultural background (25). From a practical standpoint the development of a battery of tests of the sort described here may afford an opportunity to improve the nature of individual psychological testing by the use of brief tests tapping separate aspects of a common function. Such a development has indeed been advocated by investigators like Zubin (35), who have been critical of the interpretative complexities inherent in the use of global projective devices like the Rorschach.

---

[3] A recent paper by Foster (6) which came to the attention of the authors subsequent to the completion of this study also describes a factor termed "emotional surgency" which has high loadings for Rorschach shading and color responses.

The two remaining factors which emerged in this study, while not anticipated, provide interesting possibilities for fruitful study as well. *Emotional Surgency,* with its loadings on ideational productivity, aggressive or diffuse motor behavior, and Rorschach Color and Shading, calls attention to the possibilities for exploring still another dimension of perception and action. It is possible that some element of lack of perceptual control (34) goes along with a concomitant inability to restrain motor behavior. This hypothesis would not sufficiently explain the high loading of Rorschach Animal Movement responses and of Form-dominated Color responses in this sample, however, and it remains to be seen whether further studies duplicate these results. Factor B, tentatively termed *Ambitiousness* or *Need-Achievement,* suggests the picture of individuals who feel compelled to assert that they are planful and can do well in a task they have not attempted before and press hard for such accomplishment. It may be important to note that IQ, which does not load appreciably on any of the factors, has its highest value in Factor B. Thus, at least for this sample of schizophrenics, it appears that a personality dimension involving achievement drive and a need to appear planful may be delineated and proves to be relatively independent of fantasy tendencies.

A final question that may be raised in connection with these findings of the second-order analysis is whether the four factors bear any diagnostic psychopathological significance. One might, for example, speculate on a hierarchical ordering of Factors BB and AA relative to an ideal criterion of mental health (or well-integrated ego functioning). Thus, Factor BB with its pattern of planfulness, motor control, and ward cooperativeness as well as ambitiousness, might represent a more effective level of functioning than Factor AA, where there is a suggestion of excessive withdrawal (Low External Interests) and Emotional Surgency with its indication of aggressive behavior and uncooperativeness as well as diffuse motor activity. Since duration of illness and florid symptomatology did not appear to differentiate between various subgroups in this sample, such an ordering of psychopathological tendencies seems less tenable although it remains reasonable to expect that an optimal association of fantasy and delaying capacity would lead to more effective living. For the present these results, like those of Barron (1), suggest that differing styles or patterns of fantasy and motility behavior are indeed discriminable through the use of relatively simple psychological testing procedures. It seems clear that many of the problems raised by consideration of these results are relevant for general personality theory and indicate the value of research that is not geared toward differentiating one dubious psychiatric nosological group from another, but rather toward clarifying the interrelations of certain key human functions as the basis for what may ultimately become a more meaningful and objective system of personality classification. That the delineation of functional personality dimensions is not purely speculative is suggested by the fact that Factors A, D, and C here bear some resemblance to factors reported by Foster (6) as "Delay" and "Emotional Surgency," by Cattell (4) as "Nervous, alert reactivity" and "Cultured introspective self-control," as well as by Thurstone (28) as "Reflectiveness" and "Impulsivity." The work of Block and Martin (3) in delineating "Under-" and "Over-controlling" children seems also to bear directly on these factors.

Summary and Conclusions

A number of concepts drawn from the psychoanalytic theory of thought development, Werner's sensory-tonic theory of perception, and Rorschach's work on the Human Movement response (M), have led to the postulation of a dimension of personality which involves a linkage of fantasy tendencies with control of motility and impulsive behavior. In order to test this hypothesis a battery of tests which presumably measured fantasy tendencies, level of aspiration, planning ability, delaying capacity, motor inhibition and motility, and responsiveness in ward behavior was developed. The battery was administered to 100 male schizophrenic patients. A factor analysis yielded four centroid factors which were subjected to an oblique rotation to simple structure. The four rotated factors were tentatively designated as:

1. Motor Inhibition and Planfulness (high loadings on Rorschach M, Movement-Threshold, Porteus Maze Test Quotient, Motor Inhibition Time, and Cooperativeness in ward behavior); *2. Ambitiousness or Need-Achievement* (high loadings on Initial Expectation in a level-of-aspiration situation, Initial Performance in the same task, Planning Self-Rating, Transcendence in TAT stories, and Rorschach Response Total); *3. Emotional Surgency* (high loadings on Rorschach Responses, Animal Movement, Shading and Color, TAT Transcendence, Aggressiveness on the ward, Uncooperativeness, and Diffuse Motor Activity); *4. Introspectiveness or Introversiveness* (high loadings on Rorschach M, Animal Movement Responses, Movement Threshold, TAT Transcendence, Admission of Distress following Failure, Continuation of Task in Compliance with Authority, and Lack of External Interests in Ward Behavior).

The high loadings of Movement-threshold and Rorschach's M on both the Motor Inhibition and the Introspective Factors suggest partial support for the original hypothesis but the emergence of two separate factors for motility and fantasy activity indicates that the original formulation was overly simplified. The *Emotional Surgency* factor affords some support to Piotrowski's analysis of the behavioral implications of Rorschach Shading and Animal Movement responses. Two second-order factors, one linking Emotional Surgency and lack of Introspection and the other Motor Inhibition and Ambitiousness also emerged.

Received December 8, 1955.

References

1. Barron, F. Threshold for the perception of human movement in inkblots. *J. consult. Psychol.*, 1955, **19**, 33–38.
2. Block, J., & Block, Jeanne. An interpersonal experiment on reactions to authority. *Hum. Relat.*, 1952, **5**, 91–98.
3. Block, Jeanne, & Martin, B. Predicting the behavior of children under frustration. *J. abnorm. soc. Psychol.*, 1955, **51**, 281–285.
4. Cattell, R. B. The principal replicated factors discovered in objective personality tests. *J. abnorm. soc. Psychol.*, 1955, **50**, 291–314.
5. Edwards, A. *Statistical analysis for students in psychology and education.* New York: Rinehart, 1946.
6. Foster, A. The factorial structure of the Rorschach test. *Texas Rep. Biol. Med.*, 1955, **13**, 34–61.
7. Freud, S. The interpretation of dreams. In *The standard edition of the complete psychological works*, Vol. V. London: Hogarth, 1953. Pp. 339–628.
8. Freud, S. Formulations regarding the two principles in mental functioning. In *Collected Papers*, Vol. IV. London: Hogarth, 1946. Pp. 13–21.
9. Hartmann, H. Comments on the psychoanalytic theory of the ego. In *The psychoanalytic study of the child*, Vol. V. New York: International Universities Press, 1950. Pp. 74–96.
10. Klopfer, B., & Kelley, D. M. *The Rorschach technique.* Yonkers, N. Y.: World Book Co., 1942.
11. Kris, E. The development of ego-psychology. *Samiksa*, 1951, **5**, 153–168.
12. Lorr, M. Multidimensional scale for rating psychiatric patients. Hospital Form. *VA Tech. Bull.*, Nov. 16, 1953, No. 10–507.
13. Luria, A. P. *The nature of human conflicts.* New York: Liveright, 1932.
14. Meltzoff, J., Singer, J. L., & Korchin, S. J. Motor inhibition and Rorschach movement responses: a test of the sensory-tonic theory. *J. Pers.*, 1953, **21**, 400–410.
15. Phillips, L., & Smith, J. G. *Rorschach interpretation: Advanced technique.* New York: Grune & Stratton, 1953.

16. Piotrowski, Z. A. A Rorschach compendium, revised and enlarged. *Psychiat. Quart.*, 1950, **24**, 543–596.
17. Porteus, S. D. *The Porteus Maze test and intelligence*. Palo Alto: Pacific Books, 1950.
18. Rapaport, D. *Diagnostic psychological testing: the theory, statistical evaluation, and diagnostic application of a battery of tests*. Vol. II. Chicago: Yearbook Publishers, 1946.
19. Rapaport, D. On the psychoanalytic theory of thinking. *Int. J. Psychoanal.*, 1950, **31**, 161–170.
20. Rorschach, H. *Psychodiagnostics*. New York: Grune & Stratton, 1942.
21. Schachtel, E. Projection and its relation to character attitudes and creativity in the kinesthetic responses. *Psychiatry*, 1950, **13**, 69–100.
22. Singer, J. L. Delayed gratification and ego-development: implications for clinical and experimental research. *J. consult. Psychol.*, 1955, **19**, 259–266.
23. Singer, J. L., & Spohn, H. E. Some behavioral correlates of Rorschach's experience-type. *J. consult. Psychol.*, 1954, **18**, 1–9.
24. Singer, J. L., & Herman, J. Motor and fantasy correlates of Rorschach human movement responses. *J. consult. Psychol.*, 1954, **18**, 325–331.
25. Singer, J. L., & Opler, M. K. Contrasting patterns of fantasy and motility in Irish and Italian schizophrenics. *J. abnorm. soc. Psychol.*, 1956, **53**, 42–47.
26. Small, K. Planning as a non-intellective component of intelligent behavior. Unpublished doctor's dissertation, Teachers Coll., Columbia Univer., 1954.
27. Thurstone, L. L. *Multiple-factor analysis*. Chicago: Univer. of Chicago Press, 1949.
28. Thurstone, L. L. The dimensions of temperament. *Psychometrika*, 1951, **16**, 11–20.
29. Weisskopf, Edith A. A transcendence index as a proposed measure in the TAT. *J. Psychol.*, 1945, **19**, 317–327.
30. Werner, H. Motion and motion perception: a study of vicarious functioning. *J. Psychol.*, 1945, **19**, 317–327.
31. Werner, H., & Wapner, S. Sensory-tonic field theory of perception. *J. Pers.*, 1949, **18**, 88–107.
32. Werner, H., & Wapner, S. Toward a general theory of perception. *Psychol. Rev.*, 1952, **59**, 324–333.
33. Wilensky, H. The performance of schizophrenic and normal individuals following frustration. *Psychol. Monogr.*, 1952, **66**, No. 12 (Whole No. 344).
34. Wittenborn, J. R. A factor analysis of Rorschach scoring categories. *J. consult. Psychol.*, 1950, **14**, 261–267.
35. Zubin, J. Objective evaluation of personality tests. *Amer. J. Psychiat.*, 1951, **107**, 569–576.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

The Development of a Short Form of the Manifest Anxiety Scale[1]

A. W. Bendig
University of Pittsburgh

Two studies (1, 2) have indicated that a majority of the 50 items on Taylor's Manifest Anxiety Scale (MAS) are not valid in predicting clinical criteria of manifest anxiety. The consistency in item validity results of independent studies suggests that a shortened form of the MAS, retaining only the valid items, might be more useful and clinically valid than the standard MAS. However, such a shortened MAS might lack the desirable normative characteristics, such as high reliability, of the longer MAS.

The 20 most consistently valid MAS items in the reports noted above (Item 1–16 and 29–32 in 2, p. 360) were selected as a shortened form of the MAS. The 50-item MAS was administered to 744 college *S*s and their papers scored for the 20-item scale. The 20-item scale, omitting the 30 "nonvalid" items, was given to an additional 324 *S*s. No significant differences in scale means or variances were found for the two methods of administration, nor were any sex differences evident. A survey of studies using the 50-item MAS shows its median internal consistency reliability to be .82, while the similar reliability of the 20-item scale was .76. Three scores were obtained for 100 random *S*s who had taken the 50-item form: (*a*) score on all 50 items, (*b*) score on the 20 "valid" items, and (*c*) their scores on the 30 "nonvalid" items. The reliabilities of the three scores were: (*a*) .78, (*b*) .76, and (*c*) .48. The intercorrelations among the scores were: (*a*) and (*b*), .93; (*a*) and (*c*) .91; and (*b*) and (*c*), .70.

It was concluded that the 20-item Pittsburgh revision of the MAS (*a*) has eliminated from the standard MAS items of low internal consistency and validity; (*b*) provides scores that are about as reliable as the 50-item MAS and are highly related to scores on the standard form; and (*c*) is more parsimonious of testing time and probably more valid than the longer MAS.

Brief Report.
Received June 25, 1956.

References

1. Buss, A. H. A follow-up item analysis of the Taylor Anxiety Scale. *J. clin. Psychol.*, 1955, **11**, 409–410.
2. Hoyt, D. P., & Magoon, T. M. A validation study of the Taylor Manifest Anxiety Scale. *J. clin. Psychol.*, 1954, **10**, 357–361.

---

[1] An extended report of this study may be obtained without charge from A. W. Bendig, Dept. of Psychology, University of Pittsburgh, Pittsburgh 13, Pa., or for a fee from the American Documentation Institute. Order Document No. 4990 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Perception of Apparent Motion in Organics and Schizophrenics

Rayford T. Saucer
VA Hospital, Perry Point, Maryland

and Herdis L. Deabler
Gulfport Division, VA Center, Biloxi, Mississippi

Klebanoff, Singer, and Wilensky, in their article summarizing (3) recent findings concerning the psychological consequences of brain lesions and ablations, make the following statement:

> Perceptual research with brain-injured patients suggests that the repertory of the clinical psychologist in the examination of neurological patients may soon be augmented by the use of specialized laboratory methods such as tachistoscopic presentation, critical flicker fusion frequency testing, apparent movement thresholds, and other so-called "brass instrument" techniques. The value of these perceptual techniques in localization of lesions remains equivocal. It seems clear, nevertheless, that subtle perceptual impairments or reorganizations following cerebral injury are most clearly brought out by intensive laboratory examination rather than by the widely used global clinical methods . . . (3, p. 11).

The present paper is a report of an experiment in the perception of apparent motion, together with its theoretical bases and implications, and attempts to show that as a result of increased knowledge in this area, future research in it may lead to the development of a reasonably critical diagnostic instrument.

Werner and Thuma (10) found that brain-damaged children failed to report perception of tachistoscopic apparent motion. Relating the deficit to Goldstein's theory of perception, they attributed it to inability to organize figure-ground relationships, which they held to be a fundamental principle of neurophysiological organization. Shapiro (9) attributed the deficit to satiation in the cortex, equating the effect of continuous visual stimulation to the effect of organic damage, in accordance with Köhler and Wallach's (4) theory of satiation.

Price and Deabler (6), in a study designed to extend Freeman and Josey's (1) work on perception of the apparently moving afterimage induced by a rotating Archimedes spiral, report that 98 per cent of the known organic patients tested had difficulty in perceiving or failed to report this distinctive illusion. In discussing their results, they suggest that either a field theory or a theory of perception proposed by one of the writers may be used to explain their findings.

Earlier work in the field of apparent motion by one of the writers (7, 8) has led to the development of new techniques for the study of apparent motion. In the present study, apparent motion is produced by use of two alternately flashing light bulbs which are 1¼-inch neon-filled, fluorescent coated Westinghouse Nite-Lites mounted on a vertical black board. These are illuminated alternately, with zero pause time between flashes, by a multivibrator square wave generator, the circuit of which is conventional.

It is found that varying the rate of alternation of light flashes from 2 to 10 per second causes the light stimuli to appear in three phases which are directly related to the frequency of flashes. As the frequency increases, the stimuli may be perceived and described as (*a*) successive alternating appearance of each light; (*b*) apparent movement wherein the two lights seem to fuse into one light which appears to swing smoothly from side to side; and (*c*) simultaneous appearance wherein the two lights are seen and perceived as flickering. The distinctive apparent movement phenomenon perceived in Phase *b* was observed by Wertheimer in 1912; he referred to it then as "optimal movement" (11). It has since been investigated by numerous psychologists (2, p. 899), being uniformly designated as Beta movement, by which term it will be referred to in this study.

If the light stimuli are made diffuse, with no sharply defined boundaries, such as is accomplished by use of a ground-glass screen placed about one inch in front of the bulbs, then the last state, Phase *c*, may result in a phenomenon earlier reported (7) called Omega movement. Omega movement appears as movement of a dark, quasi-formless, amorphous object. It has movement but differs from Beta in that it does not have the attributes of the primary stimulus. It has a minimal amount of "thing" quality and is frequently reported as "a dark shadow moving back and forth."

The body of evidence indicates a deficit in organics in their ability to perceive apparent motion. It is felt that the deficit reported by Price and Deabler may be a phase of a general disturbance of the perception of apparent motion. The present study is designed to explore further the nature of this deficit.

A theory of motion perception has been previously presented (7, 8). This theory states that motion perception is based on a matrix synthesis of perceptual organizations isomorphic or related to stimulus detail and to temporal aspects of the stimulus configuration. Beta motion, in which the apparently moving object has the attributes of the primary stimulus, may be regarded as a synthesis of stimulus detail and motion. When the rate of alternation is increased from that which gives optimal Beta motion, the complex or synthesis breaks apart or dissociates. Perception of Beta movement is lost and the observer reports seeing either the two lights as stationary or the Omega movement phenomenon. Practically speaking, the threshold for loss of Beta is the same as for appearance of Omega. It is felt that the ability to sustain Beta perception at higher levels of complexity is dependent upon the functional efficiency of the cerebral cortex. The threshold for loss of Beta perception may thus lend itself as a measuring index for determining cortical efficiency or impairment.

It is to be expected that organics with their decreased cortical efficiency will have lower thresholds for loss of Beta or for appearance of Omega than normals. Schizophrenics by operational definition and by their performance on most psychological tests are known to have low cortical efficiency also. They are included in this study to determine whether or not this is reflected in lower Omega threshold or whether this psychiatric group remains essentially normal with respect to ability to perceive apparent motion.

It is the specific hypothesis of this study that as the rate of alternation of the two lights is changed, organics will have a lower experimental threshold for Omega or loss of Beta than normals, and it is highly probable that schizophrenics will also be lower than normals.

Subjects and Method

Four groups of 10 male subjects per group were used in this experiment, three of the groups being drawn from the patient population of a psychiatric hospital and one group from the population of nursing aides and attendants. The groups were designated as organic, schizophrenic, lobotomized, and control.

Diagnosis of patients by the hospital staff was used as the criterion for grouping. The official diagnosis for each patient represented the opinion of psychiatrists, psychologists, and occasionally, psychiatric consultants. Patients were selected from the current case book of the Psychology Service and represented a random sampling of a hospital population. It is believed that they were typical for their particular diagnostic groups.

Although many organic subjects had a history of one or more brain injuries, no particular category of organic involvement was favored. Alcoholic and senile syndromes were excluded. Practically all organic patients tested carried a secondary or qualifying diagnosis of "with psychotic reaction."

Schizophrenics were taken from the category of schizophrenic reaction, undifferentiated type. If a history of concussion damage, blackouts, or other disorders of consciousness was elicited, subjects were not used. All schizophrenic patients used in the experiment had experienced delusions or hallucinations at some period during their psychosis. Patients with a history of ECT or ICT during the previous three months were not used.

Ten Freeman-Watts lobotomy cases were also included in the study for comparative purposes. All were diagnosed chronic schizophrenics prior to the lobotomy.

The controls were selected at random from male aides employed in the psychiatric hospital. They were used because it was felt that they represented a group of individuals of average intelligence and similar cultural background to the majority of the patients used.

Subjects were brought into a darkened room and seated two meters from the light source. In a pilot study, it was found that a flash rate of 2.5 flashes per second gave optimal Beta movement for most subjects. This rate was then used as a baseline for the experiment. The rate control was set at 2.5 f.p.s., and each subject was asked to report what he saw. Of the 40 subjects tested, two organic subjects failed to report Beta movement; of these, one reported Omega and another reported seeing no motion of any description. The flash rate was then raised in steps of 0.5 f.p.s. every 30 seconds until the subject reported Omega movement. In a few cases the Beta organization was broken up by simultaneous appearance of two fused lights, which represented a high threshold of dissociation. This point was equated to the perception of Omega motion (8).

Results

The rate at which Omega motion became visible or the rate at which the two lights appeared simultaneously was taken as the threshold of dissociation of Beta movement. The mean rates for the diagnostic groups and their associated standard deviations are given in Table 1.

It can be seen from Table 1 that the means can be ranked from organic through normal. Analysis of the significance of the difference of the means is contingent upon the results of the analysis of variance and will be presented in that order.

Table 1

Threshold Means and Standard Deviations of Groups

Measure	Organics	Schizophrenics	Lobotomies	Control
Means	3.58	3.96	4.45	4.52
Sigmas	.22	.13	.22	.14

Table 2

Analysis of Variance

Source	Sum of Squares	*df*	*F*
Total	19.19	39	
Between	5.94	3	5.50
Within	13.26	36	

Since a test for homogeneity of variance indicated that the variance of the groups represents no more than chance distribution, analysis of variance for between and within categories was applied to the data. The results are given in Table 2.

For 3 and 36 degrees of freedom, an F ratio of 5.50 is significant beyond the .01 level but not to the .001 level of significance. The null hypothesis, that these groups represent random samples drawn from a common population, can be rejected at a satisfactory level of confidence. It may then be concluded that the threshold for the dissociation of Beta apparent motion is related to the presence of pathology.

Following McNemar (5, pp. 74–75), the standard error of the difference can be computed by using the within-variance as the best estimate of the population variance. If this is done, the standard error of the difference for the groups can be based on $N - k$ degrees of freedom, where N is 40 and k is 4. Thus, 36 degrees of freedom are available for the various tests. The various tests are summarized in Table 3.

Table 3

Significance of Mean Differences

Groups Compared	*t*	Significance
Organic-Schizophrenic	1.5	NS
Organic-Lobotomy	3.3	.01
Organic-Normal	3.5	.01
Schizophrenic-Lobotomy	1.7	.05
Schizophrenic-Normal	2.1	.025
Lobotomy-Normal		NS
OS-LN	3.78	.01
OS-N	3.40	.01

For four of the pairs, a *t* which is not quite significant at the .001 level of significance but well beyond the .01 level can be established. Two are not significant and two are significant at the .05 level of confidence. Current statistical practice sets no level for rejecting the null hypothesis, although the writers feel that the confidence level should at least exceed .01 before the null hypothesis is rejected. If we wish to consider these ratios as a *CR* (since 36 degrees of freedom are involved), then those ratios above 3.3 rise considerably above the .001 level of significance. The writers prefer the more conservative evaluation.

Discussion

The primary hypothesis that organics differ from normal or control subjects in Omega threshold has been established at a highly significant level of significance. Thus we may expect that as test procedure and instrumentation are refined, a diagnostic instrument may be evolved which will provide a statement of degree of pathology but not necessarily indicate the kind. Although schizophrenics are not distinguishable from organics on the basis of this technique, the distinction between normals and schizophrenics suggests that an apparent-motion perception technique may assist in differentiating these conditions.

At the same time, significant *F* ratio indicates a significant nonchance order of mean scores. If the *F* ratio be taken as a limited case of a correlation coefficient, the Omega threshold is then correlated with some unknown function. This function is not necessarily the degree of behavioral deviation exhibited by the patient. The writers believe that the means represent points on a continuum of cortical efficiency.

If the results are examined in the light of a neural network frame of reference, they become meaningful. A matrix theory of perception can be expanded to include the entire cerebral cortex as a single matrix. Any damage to any part of this matrix represents a loss and hence lowers the functional efficiency of the matrix. One of the major functions of the cortical matrix is to mediate contact between the organism and the environment by ordering the information received through sensory channels. The subjects selected as organic demonstrate damage to the visual perceptive mechanisms regardless of the type of cortical damage. Evidently perception cannot be fractioned into visual or other components but is a global effort of the entire cortex. The continuum on which the subjects are ranked is the continuum of over-all efficiency of the cerebral cortex, not the efficiency of the visual pathways and visual projection areas alone.

The organic subject exhibits deficit because of damage to neural processes. The schizophrenic, on the other hand, shows deficit because of psychological factors which may or may not have a demonstrable isomorphic physiological basis. A large part of the cortex of the schizophrenic is involved in dealing with problems arising from emotions such as hostility toward a basically unfriendly environment. Neural matrices normally used in mediating environmental contact are disrupted by being organized into matrices concerned with problems which we may broadly define as emotional or as internal. Such neural mechanisms as are involved in these autistic processes must be subtracted from the sum-total amount of cortical tissue available to the organism for the mediation of contact with the environment. By a law of mass action, the results are not to be clearly separated from the results of organic damage.

The interpretation of schizophrenic responses may be inferred from the fact that when lobotomized patients are tested they fall toward the "normal" rather than the "deficit" end of the continuum. As has been mentioned, these patients were chronic schizophrenics unimproved by other therapies. When the subcortical tracts which mediate emotional responses to sensory material are out, they tend to return to the norm with respect to perceptual organization. As the tissue damage is mainly in subcortical areas, the functional Gestalt of the cortex is not destroyed.

From the experimental evidence, it may be inferred that perception is a global process of the entire cortex and that by measuring the amount of organizational force available to the individual for use in perceptual processes such as perception of apparent motion, the functional efficiency of the individual may be measured and presence or absence of pathology determined.

Summary

Experiments reported by previous investigators indicated a deficit in the perception of apparent motion by organic individuals. The hypothesis that the deficit is due to lowered organizational ability has been advanced. The degree of organizational ability of the individual would be reflected in the threshold for dissociation of Beta apparent motion into component parts of detail and Omega movement.

Organics, schizophrenics, lobotomy cases, and controls were examined for threshold of Beta dissociation. Analysis of variance indicates that the null hypothesis—that these four groups are random samples of a common population with respect to the variable measured—can be rejected at beyond the .01 level of significance.

It is concluded that the threshold for loss of Beta apparent motion can differentiate between certain clinical groups, and it is inferred that these differences are due to organic or functional loss of organizing capacity of the cerebral cortex as a whole rendering it less able to synthesize sensory data.

Received November 1, 1955.

References

1. Freeman, E., & Josey, W. E. Quantitative visual index to memory impairment. *Arch. Neurol. Psychiat.*, 1949, **62**, 794–796.
2. Graham, C. H. Visual perception. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 868–920.
3. Klebanoff, S., Singer, J. L., & Wilensky, H. The psychological consequences of brain lesions and ablations. *Psychol. Bull.*, 1954, **51**, 1–41.
4. Köhler, W., & Wallach, H. Figural aftereffects; an investigation of visual processes. *Proc. Amer. phil. Soc.*, 1944, **88**, 269–357.
5. McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.
6. Price, A. C., & Deabler, H. L. Diagnosis of organicity by means of spiral aftereffect. *J. consult. Psychol.*, 1955, **19**, 299–302.
7. Saucer, R. T. Nature of perceptual process. *Science*, 1953, **117**, 556–558.
8. Saucer, R. T. Processes of motion perception. *Science*, 1954, **120**, 806.
9. Shapiro, M. B. A preliminary investigation of the effects of continuous stimulation on the perception of "apparent motion." *Brit. J. Psychol.*, 1954, **45**, 58–67.
10. Werner, H., & Thuma, A. A deficiency in the perception of apparent motion in children with brain injury. *Amer. J. Psychol.*, 1942, **55**, 58–67.
11. Wertheimer, M. Experimentelle Studien über das Sehen von Bewegung. *Z. Psychol.*, 1912, **61**, 161–165.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Emotional and Intellectual Concomitants of Advanced Chronic Alcoholism

George A. Peters, Jr.[1, 2]
Boston Regional Office, Veterans Administration

Much has been written concerning the effects of excessive drinking on human mental processes. The chronic alcoholic has been variously labeled as psychotic, psychoneurotic, psychopathic, and intellectually deteriorated. However, these beliefs have been based upon studies which have dealt with obviously biased samples drawn from such sources as alcoholics' courts, mental hospitals, psychoanalysts' case loads, and the prisons.

A sample thought to be more representative of chronic alcoholics in the general population was drawn from inpatients on the alcoholism ward of a large Eastern general hospital. A group of 120 males, virtually all the chronic alcoholics admitted for treatment during the prescribed period of this study, were examined. The source of patients was quite varied and there were many case histories of delerium tremens and visual and auditory hallucinosis associated with the use of alcohol. These patients had been drinking excessively for from 10 to 30 years, were acutely alcoholic shortly before admission, and were physically debilitated as a result of prolonged alcoholic excess.

Administration of the Wechsler-Bellevue Intelligence Scale revealed a wide range of mental ability, little variation between the verbal and performance subtest scores, and a slightly higher mean score than found in the general population standardization data. (Full Scale mean 106.2, *SD* 13.89; Verbal mean 106.1, *SD* 14.21; Performance mean 105.1, *SD* 13.72.) Academic accomplishments ranged from below the level of basic education to possession of a doctoral degree. The median school grade completed was the tenth grade.

Diagnosis in accordance with the psychobiologic nomenclature of the American Psychiatric Association indicated that these chronic alcoholics were not primarily psychotic (8.3%), psychoneurotic (7.5%), nor sociopathic (5.0%). The number classified as chronic or acute brain syndrome (5.8%) indicates that only a small minority developed clinical symptoms of central nervous system "organicity." Over half (52.5%) were classified as manifesting symptoms indicative of a Personality Disorder.

This group of patients suffering from the physical complications of advanced chronic alcoholism displayed *no* evidence of marked intellectual deterioration on the basis of gross IQ scores. The majority of cases were best described as manifesting a variety of personality disorders rather than psychoses, neuroses, sociopathic, or brain syndromes. If this sample is used as a criterion, there is a surprising lack of major psychiatric disorder or restriction of intellectual capability in the advanced stages of chronic alcoholism.

Brief Report.
Received April 2, 1956.

---

[1] Formerly Senior Psychologist, Alcoholism Study Unit, Philadelphia General Hospital, Philadelphia, Pa. Gratitude is expressed for the aid given by Drs. Martin D. Kissen, Thomas Downs, and Albert Biele of the Philadelphia General Hospital.

[2] An extended report of this study may be obtained from George A. Peters, Jr., 65 Manor Drive, Newark 6, New Jersey or for a fee from the American Documentation Institute. Order Document No. 4971 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.75 for microfilm or $2.50 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Repression and Insight as Related to Reaction to Cartoons

Seymour Epstein and Richard Smith
University of Massachusetts

A commonly accepted theory of humor is the Freudian view that repression favors an appreciation of humor in which the repressed material is represented (7). Although several studies on the nature of the humor response have been reported, most of which have endorsed Freud's view, a critical appraisal of these studies indicates that they fail to deal directly with the Freudian hypothesis. What they do reveal is that (*a*) individuals show a consistent preference for certain types of humor (1, 3, 4, 5, 8, 12, 13), (*b*) preference for certain types of humor can be related to other personality variables (2, 4, 5, 9), and (*c*) preference for certain types of humor can be related to adjustment (12).

In order to evaluate the Freudian hypothesis it would be desirable to obtain a measure of preference for a certain type of humor, to secure an independent measure of repression, and to determine whether there is a significant relationship between the two. This is what the present study proposes to do.

Method

Hostility was selected as the variable to be investigated because, as a socially disapproved trait, it is subject to repression and there is a considerable amount of humor devoted to it. In order to obtain a measure of preference for hostile humor, 16 hostile and 16 nonhostile cartoons were obtained by selecting only those unanimously agreed upon by independent decisions from five judges. The cartoons finally selected were assigned random numbers and presented to each *S* individually with two sets of instructions. First *S* was to rate each cartoon by placing it into one of three bins according to whether it was "funny," "mediocre," or "not funny." As *S* observed each cartoon, *E* (examiner) unobtrusively recorded on a checklist whether *S*'s expressive reaction was: (*a*) a negative verbal expression, (*b*) a negative facial expression, (*c*) no reaction, (*d*) a positive facial expression, (*e*) a positive verbal expression. After *S* had completed sorting the cartoons, *E* shuffled them and instructed *S* to rearrange them in a Q sort by placing the funniest cartoons farthest to the right and the least funny farthest to the left. The Q sort contained seven columns with successive frequencies 2, 4, 6, 8, 6, 4, 2. The Q sort differed from the initial ratings in that it required that each cartoon be judged relative to all other cartoons.

Establishing a measure of repression was more difficult than measuring reaction to humor. According to Freud, "The essence of repression lies simply in the function of rejecting and keeping something out of consciousness" (7, p. 86). Accordingly, it would seem reasonable to assume that repressed impulses would achieve expression in everyday behavior to a greater extent than they would be consciously acknowledged. Consequently, an operational measure of repressed hostility was established as the degree to which a person denied the hostility attributed to him by a consensus of persons who knew him over an extended period. A measure of insight was obtained by considering the absolute discrepancy without regard to direction (2, 11).

Thirty-two fraternity members who had known each other for at least a year served as *S*s. A week after having sorted the cartoons, each *S* was again seen individually and presented with 32 cards containing the names of all *S*s, including his own. It was requested that the cards be placed in a Q sort such that the person who most approached the following description would be farthest to the right: "This person is an individual who tends to be inconsiderate, antagonistic, and resentful. When talking about others he is apt to emphasize their faults although he does not like this done to himself. He has little sympathy for others and enjoys putting them on the spot. He makes disparaging remarks about people and argues for the sake of argument. It is important for him to come out on top and he is apt to be a sore loser in competition." This definition was made to be relevant to the manifestations of hostility that were apt to occur in group living. The Q sort consisted of seven columns with successive frequencies of 2, 4, 6, 8, 6, 4, 2.

Results

An index of repression and insight was obtained by subtracting from the self-rating of hostility on the Q sort, the average rating by others. Inspection revealed that there were nine individuals whose self-ratings were higher than their average rating by others. The *S*s, therefore, were divided into three groups of nine each: (*a*) a group that overestimated its hostility and was moderately lacking in insight, with a mean self-minus-others score of 1.03, (*b*) a group that very slightly underestimated its hostility and was relatively insightful, with a mean of − 0.36, and (*c*) a group that considerably underestimated its hostility and was most lacking in insight, with a mean of − 1.72. For simplicity, the groups will hereafter be referred to as Repression Groups and attention will be turned to insight only when necessary. Differences between the groups were evaluated by analysis of variance of the measures of expressive reaction, cartoon rating, and cartoon Q sort. As there was some danger that the measures of insight and repression were confounded with absolute level of hostility, the groups were compared on mean hostility ratings by a simple between-within analysis of variance. An F of less than one was obtained indicating that any differences in over-all hostility were well within chance expectancy.

In scoring expressive reactions, a weight of one was given to facial expression and a weight of two to verbal expression. An analysis of variance was then performed which allowed for an evaluation of the following: Repression Groups (R), Subjects in same group (S), Favorable vs. unfavorable reaction (F), Hostile vs. control cartoons (C), and the interactions $R \times F$, $R \times C$, $R \times F \times C$, $S \times F$, and $S \times C$. In terms of the hypothesis under investigation, the critical source of variance is $R \times F \times C$, which indicates whether Repression Groups differ in favorableness of response to hostile relative to control cartoons. This interaction did not approach significance, yielding an F of less than one. A second source of variance concerning Repression Groups is the interaction $R \times C$, which indicates whether the groups differ in amount of reaction, irrespective of kind, to hostile relative to control cartoons. Here, too, the F was less than one, suggesting a lack of relationship. The only source of variance relating to Repression Groups which approached significance was R, which indicates total amount of expressive reaction. The F of 2.68 was below the 3.40 required for significance at the .05 level, but did suggest a trend which is borne out by further findings. The order of the means was 9.69 for the group that moderately overestimated its hostility, 4.94 for the group that slightly underestimated its hostility, and 5.11 for the group that considerably underestimated its hostility. Considering that the last two means are almost identical, it appears that the group that overestimated its hostility tended to be more expressive than the other two. Additional findings from the analysis of variance are as follows: (*a*) There were highly reliable individual differences (.001 level) both in amount and favorableness of expression, (*b*) Favorable expression occurred significantly more often than unfavorable expression (.001 level), and (*c*) The hostile cartoons elicited more favorable reactions than the control cartoons (.01 level).

The data from *S*'s ratings of the cartoons were analyzed in the same manner as expressive reactions. Favorable and unfavorable ratings were summated independently, the "mediocre" rating receiving a weight of zero.

Repression Groups were found to differ significantly (.01 level) on total amount of reaction in a manner confirming the nonsignificant tendency noted for expressive behavior. The only other significant source of variance (.001 level) indicated that the hostile cartoons were rated more favorably than the control cartoons.

The data from the Q sort were evaluated by assigning weights of one to seven to the successive columns into which the 32 cartoons had been sorted. Scores consisted of the weight received on each of the 16 hostile cartoons. Analysis of variance revealed that Repression Groups did not differ significantly either in their reaction to the pooled hostile cartoons or in their evaluation of the individual hostile cartoons. The only statistically significant finding (.001 level) was that some of the hostile cartoons were judged funnier than others. Despite the lack of a reliable relationship between repression and reaction to individual cartoons, the data were inspected for leads for further work. It was observed that the cartoon which was ranked first in funniness by the group that considerably underestimated its hostility was ranked sixth by the group that slightly understimated its hostility and ninth by the group that overestimated its hostility. Thus, preference for this cartoon was directly related to denial of hostility. The cartoon was one by Virgil Partch depicting a wife plunging a revolving egg-beater into the back of her husband, and saying, "Stick 'em up!" It was *E*'s impression, confirmed by independent decisions from three other judges, that this cartoon was the most straightforwardly hostile in the sample. Unfortunately, only two similar cartoons of the "practical joke" variety depicting an unprovoked act of hostility by an adult could be found. In both cases a direct relationship was indicated between repression and preference for the cartoon.

In order to determine whether it is necessary to take into account level of funniness of the cartoons, two separate analyses of variance of Q-sort ratings were done, one based on the four least funny, the other on the four most funny hostile and control cartoons as determined by the mean rating of the cartoon. On the least funny cartoons Repression Groups did not differ in preference for hostile relative to control cartoons but did differ in over-all rating of the pooled cartoons (.05 level). The order of the means was 29.33 for the group that considerably underestimated its hostility, 25.11 for the accurate group that very slightly underestimated its hostility, and 26.00 for the group that considerably overestimated its hostility. Thus, the group that evaluated itself most accurately judged the poor cartoons most appropriately, and the reverse holds for the group that evaluated itself most inaccurately. On the most funny cartoon, Repression Groups did not differ significantly in their reaction to the pooled cartoons, but the order of the means for the same three groups was, respectively, 36.44, 39.56, and 37.00, which again suggested that accuracy of self-evaluation was directly related to appropriateness of reaction to the cartoons. Thus, taken together the findings on the least and most funny cartoons indicated a direct relationship between insight and sense of humor.

In order to determine the relationship between the three measures of reaction to hostile cartoons, Pearson product-moment correlations were computed. To do this, the measure of expressive reaction was expressed as a ratio of favorable reactions to hostile cartoons over favorable reactions to all cartoons. The same index was used for cartoon-ratings. For the Q sort, the score was simply the total weight on the hostile cartoons. The following correlations were obtained: expressive reaction and cartoon-rating, .57; expressive reaction and Q sort, .52; and cartoon-rating and Q sort, .79. All correlations are significant beyond the .01 level (t test, $N = 32$). These results indicate a significant degree of reliability which is at least .52 for expressive reaction and .79 for cartoon-rating and for Q sort. That the measure of expressive reaction correlated lowest with the other two may indicate that it has the lowest reliability or that it measured something unique, or both. There is some evidence in support of the latter possibility in that expressive reaction was the only measure on which significant individual differences were found.

Discussion

Repression and humor. The results of this study fail to support the hypothesis that repression of hostility is related to a general preference for hostile cartoons. Negative results are difficult to interpret as there is no way of knowing whether different operational measures or different populations would have produced different findings. In defense of the measure of reaction to cartoons used, adequate reliability for group differentiation was indicated. Moreover, validity was somewhat supported by the finding that reaction to cartoons was meaningfully related to insight. The measure of repression is somewhat more questionable, but it, too, was meaningfully related to another variable in that it was inversely associated with over-all expressiveness. This relationship reveals a degree of reliability beyond chance, although there is no indication of how high it is. However, since the Q technique for measuring repression required each *S* to judge and be judged by all other *S*s relative to each other, an acceptable degree of reliability would be expected, at least approaching the minimum estimate of .79 obtained from the same Q sort for cartoons. Considering that all *F* values on the interaction of repression and preference for hostile cartoons were much less than one, there is no reason to suspect that higher reliability would have altered the findings. That another operational measure of repression might have produced positive findings remains to be demonstrated. Some question might be raised as to whether a group that overestimates its hostility should be considered as low in repression. If only the accurate group and the group that underrated its hostility are compared, the results are equally negative.

One explanation of the negative findings on repression and humor is that there is no *general* relationship between repressed hostility and preference for hostile cartoons. That there might be a relationship between repressed hostility and a preference for certain types of hostile cartoons was suggested by the results on a cartoon of the practical-joke type which depicted an unprovoked act of hostility. However, the evidence here was inconclusive and verification is required. A second possibility is that a relationship would be found if suppression (conscious inhibition) or external frustration of an impulse were investigated rather than repression (unconscious inhibition). This, too, remains to be determined, but seems likely in view of common observations of the interest in sexual jokes of sexually frustrated groups such as adolescents (culturally frustrated) and men in the armed services (geographically frustrated), who are not necessarily repressed. A third possibility for investigation is that a relationship between repression and humor becomes manifest only when repression reaches pathological proportions (10).

Insight and humor. Allport states, ". . . the most striking correlate of insight is the *sense of humor*" (2, p. 222). He cites one study in which a correlation of .88 was found between ratings on insight and ratings on sense of humor. In the present study it was found that there is a statistically significant relationship between rating the cartoons correctly and judging one's hostility accurately. Why are insight and humor related? Although there was no way of testing this in the present study, it is hypothesized that the two are related only when humor involves an ability to accept the self as its object. Presumably the person who is able to laugh at his limitations should be low in defensiveness and accordingly should be insightful. It would be interesting to develop a measure of differential appreciation of humor in which the object of some of the jokes or cartoons is self-related and of others is other-related. This measure might then be related to measures of defensiveness and insight. Perhaps a subtle clinical instrument for determining defensiveness and insight would result.

An incidental finding of some interest was that there are individual differences in degree and type of expressive reaction irrespective of the content of the cartoon responded to. This finding indicates that it is necessary to use control cartoons in evaluating differences in expressive reaction to any particular type of cartoon. It further indicates that individual differences in expressive behavior, apart from the nature of the cartoons responded to, is itself a promising personality variable for investigation.

Summary

Preference for hostile humor was measured by three techniques, which included observing *S*'s expressive reactions to hostile and control cartoons, having *S* rate the cartoons, and having him sort them on a Q sort of funniness. In order to obtain a measure of repression, the *S*s, members of the same fraternity for at least a year, rated themselves and each other on a Q sort of hostility. The degree to which the self-rating was lower than the average rating by others was taken as an index of repression; the absolute discrepancy between the self-rating and rating by others was taken as an index of insight. Three groups of nine each were selected, a group that considerably overestimated its hostility, an accurate group that very slightly underestimated its hostility, and a group that considerably underestimated its hostility. Following are the major findings:

1. No relationship was found between repression and general preference for hostile cartoons, but there was questionable evidence that repression was related to preference for a practical-joke type of cartoon depicting an unprovoked act of hostility.

2. A significant positive relationship between insight and sense of humor was found.

3. A significant negative relationship between repression and degree of reactivity was found.

4. Highly reliable individual differences in degree and kind of expressive reaction irrespective of the cartoons responded to were found.

Implications for further work on the relationship of humor to repression and insight were discussed.

Received January 9, 1956.

References

1. Andrews, T. G. A factorial analysis of responses to the comic as a study in personality. *J. gen. Psychol.*, 1943, **28**, 209–224.
2. Allport, G. W. *Personality, a psychological interpretation*. New York: Holt, 1937.
3. Cattell, R. B., & Luborsky, L. B. Personality factors in responses to humor. *J. abnorm. soc. Psychol.*, 1947, **42**, 402–421.
4. Eysenck, H. J. The appreciation of humor: an experimental and theoretical study. *Brit. J. Psychol.*, 1942, **32**, 295–309.
5. Eysenck, H. J. Experimental analysis of five tests of "appreciation of humor." *Educ. psychol. Measmt*, 1943, **3**, 191–214.
6. Freud, S. *The basic writings of Sigmund Freud*. New York: Random House, 1938.
7. Freud, S. *Collected papers*. Vol. 4. London: Hogarth, 1950.
8. Luborsky, L. B., & Cattell, R. B. The validation of personality factors in humor. *J. Pers.*, 1947, **15**, 283–291.
9. Murray, H. A., Jr. The psychology of humor. *J. abnorm. soc. Psychol.*, 1934, **29**, 66–81.
10. Redlich, F. C., Levine, J., & Sohler, T. P. A mirth response test; preliminary report on a psychodiagnostic technique utilizing dynamics of humor. *Amer. J. Orthopsychiat.*, 1951, **21**, 717–734.
11. Sears, R. R. Experimental studies of projection. In S. S. Tomkins (Ed.), *Contemporary psychopathology*. Cambridge, Mass.: Harvard Univer. Press, 1943.
12. Strother, G. B., Barnett, Margaret M., & Apostolakos, P. C. The use of cartoons as a projective device. *J. clin. Psychol.*, 1954, **10**, 38–42.
13. Yarnold, J. K., & Berkeley, M. H. An analysis of the Cattell-Luborsky Humor Test into homogeneous scales. *J. abnorm. soc. Psychol.*, 1954, **49**, 543–546.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

The Meaning and Generality of Perceived Actual-Ideal Discrepancies

Leon H. Levy[1]
Indiana University

Recent studies by Rogers (4) and his associates have revealed that when subjects (*Ss*) are asked to do Q sorts of self-referent items under instructions to sort them so that they describe themselves as they are at present and then again so that they describe their concept of the ideal person, the correlation between these two sorts seems to vary with the presence or absence, or severity, of maladjustment. The size of the correlation between actual- and ideal-self-sorts has been interpreted variously as a measure of self-esteem or self-acceptance, and the finding that persons classified as maladjusted have lower Q correlations than persons not so classified has been taken as evidence for the importance of self-acceptance in adjustment and the value of the self concept as a construct in personality theory. Other investigators have suggested defining personal maladjustment "as any discrepancy between the concept of self and the concept of ideal self" (1, p. 257).

However, upon careful analysis of the procedures and data involved in these experiments it does not appear that these inferences are entirely justified. In this paper we propose to examine the logical justification for these inferences and then to present data supporting an alternate hypothesis.

The task set for the *S* in the typical actual-ideal Q-sort study is one of describing one aspect of his environment as he perceives it in its actual state and again as he perceives its ideal state to be. The resulting correlation may be interpreted as an indication of the perceived discrepancy between the actual and ideal states of the particular aspect of the environment delineated for the *S* by the experimenter's instructions and the descriptive items he has to work with. In the studies referred to above the aspect of the environment involved is labeled "the self" and a set of verbal items deemed by the experimenter as appropriate in describing the self function to delineate that aspect of his total environment for which the *S* is to indicate his perceived actual-ideal discrepancy.

The finding that a certain portion of the variance of the discrepancy measure may be attributable to individual differences does not in and of itself permit any inferences concerning the significance of the particular aspect of the environment used in obtaining this discrepancy measure. That is, until it has been demonstrated that actual-ideal correlations vary with adjustment if and only if they are obtained through the use of the self as the object of description, inferences concerning the importance of the nature of the self concept in personality functioning are not justified. It may be, for example, that the more maladjusted the individual the greater the discrepancy he perceives between the actual and ideal state of affairs over a wide range of areas of his environment. Thus, conceivably, one might discover a significant relationship between Q correlations for actual-ideal descriptions of geographical climate and personality adjustment. Unless one went on to sample other aspects of the *S*'s environment as objects of description he might be led to construct a theory of personality functioning with climatological perception as a central construct.

---

[1] The author wishes to thank Mr. Richard F. Gundy who ran the *Ss* and performed the statistical computations.

The hypothesis which we propose to test in the present study is that individuals differ in the extent to which they perceive discrepancies between the actual and ideal states of various aspects of their environment and that these differences are fairly constant from one area to another. Thus an individual obtaining a high Q correlation when doing actual-ideal sorts of self-referent items would be expected to obtain a high Q correlation when doing actual-ideal sorts of items descriptive of other aspects of his environment, and similarly with an individual obtaining a low Q correlation between actual-ideal self-sorts.

Procedure

The *S*s used in this experiment consisted of 21 volunteers (2 females and 19 males) from an introductory psychology course, none of whom were aware of the nature of the experiment at the time of their volunteering for it.

Each *S* was seen individually and was asked to do actual-ideal sorts of 100 self-referent items and of 100 home-town-referent items. The self-referent items were those developed by Butler and Haigh (2).[2] The home-town-referent items were collected in the following manner. Twenty students in the author's personality class were given the following instructions: "(*a*) List five things that you like most about your home town. (*b*) List five things that you like least about your home town. (*c*) List five things which you feel would be true of the ideal home town." From the resulting pool of 300 items, 100 were selected which appeared to have fairly wide applicability and which provided an approximately equal representation of positive and negative statements.

The *S*s were required to sort the statements for each sort into nine piles with the following distribution of cards per pile: 1, 4, 11, 21, 26, 21, 11, 4, 1. Each *S* was first presented with the self-referent statements and asked to sort them so as to describe himself as he sees himself at present from those least like himself to those most like himself. Following this he was asked to sort the home-town-referent statements so as to describe his own home town. Next he was asked to sort the self-referent statements again so as to describe his ideal person, or the person he would most like to be. And finally, he was asked to sort the home-town-referent statements so as to describe his concept of the ideal home town, the kind of town he would most like to live in. The same order of presentation was used with all *S*s.

The two Q correlations resulting for each *S* were then correlated over all *S*s in order to determine the extent to which actual-ideal discrepancies in the two areas sampled were related.

Results and Discussion

The median Q correlation for actual-ideal self-descriptions was .47 with a range from .08 to .82. The median Q correlation for actual-ideal home-town descriptions was .24 with a range from − .28 to .92. The Pearson correlation between the two sets of Q correlations was .70 which is significant at the .01 level. This finding makes the null hypothesis untenable and gives some support to the experimental hypothesis that the Q correlation obtained for an individual on actual-ideal self-descriptions represents a tendency to perceive actual-ideal discrepancies which transcends the particular aspect of the *S*'s environment used as the object of description. Furthermore, in the light of the present data, previous hypotheses concerning actual-ideal discrepancies which treated them as phenomena contingent upon such conditions as self-acceptance, self-esteem, self-satisfaction, etc., appear to be too narrowly conceived.

While the self concept as an important construct may still be salvaged by invoking the processes of projection and rationalization on the part of the *S* to account for generalized perceived actual-ideal discrepancies, particularly in the case of maladjusted individuals, such reasoning must be recognized as *ad hoc*.

In the present study only two aspects of the *S*s' environment were used as objects of description. Before any definite statement can be made about the generality of perceived actual-ideal discrepancies many more aspects would need to be sampled. It would appear however that research might be profitably directed toward a study of the nature and extent of perceived actual-ideal discrepancies.

---

[2] The author wishes to thank Dr. Rosalind F. Dymond for making these items available for this study.

Certain rather interesting research problems become opened up by such an approach and it may be of some value to sketch these briefly. One such problem would be: Given a certain number of objects of description, do individuals vary in the consistency of their perceived actual-ideal discrepancy over the range of these objects? What factors might best account for the variance in this consistency? What factors, if any, appear to be a function of the degree of perceived actual-ideal discrepancy in any given area of description? Do these vary from area to area? The theoretical implications of such research are difficult to anticipate but it seems fairly clear that current self-oriented neophenomenological personality theories (3, 5) would not be adequate to account for these data.

Summary

The logical justification for current interpretations of actual-ideal self-description discrepancies was called into question. An experiment was performed in which *S*s did actual-ideal Q sorts using self-referent items and using home-town-referent items. The resulting sets of correlations were found to have a statistically significant correlation of .70 between them and this was taken as evidence indicating a certain amount of generality to perceived actual-ideal discrepancies and further substantiating the criticism of current interpretations of these discrepancies when limited to the self as the object of description. Implications of these findings for future research were sketched briefly.

Received December 5, 1955.

References

1. Bills, R. E., Vance, E. L., & McLean, O. S. An index of adjustment and values. *J. consult. Psychol.*, 1951, **15**, 257–261.
2. Butler, J. M., & Haigh, G. V. Changes in the relation between self-concepts and ideal concepts. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago: Univer. of Chicago Press, 1954. Pp. 55–75.
3. Rogers, C. R. *Client-centered therapy*. Boston: Houghton Mifflin, 1951.
4. Rogers, C. R., & Dymond, Rosalind F. (Eds.) *Psychotherapy and personality change*. Chicago: Univer. of Chicago Press, 1954.
5. Snygg, D., & Combs, A. W. *Individual behavior*. New York: Harper, 1949.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Learning in the Immediate Post-Electroshock Period[1]

George A. Zirkle
Hanover College

The view has been current for some years that the period soon after electroconvulsive shock offers unique opportunities for psychotherapy with mental patients. Though many clinicians have used the period rather extensively for this purpose, others have questioned its value. For instance, Kalinowsky and Hoch write, "In our experience, psychotherapy during convulsive therapy has often been impossible because of the patient's confusion" (3, p. 118).

How valuable is the postshock period for psychotherapeutic purposes? If it is valuable, how soon after shock should psychotherapy be instituted? Answers to these questions will depend, among other things, on an analysis of the mental state of the patient after disorganizing shock, and a determination of the time pattern by which abilities are recovered. A number of research efforts have been directed toward this end, most of which have concentrated on memory functioning. Certain investigators (1, 5, 6) have reported a decided interference of shock with memory. Experimental testing has usually been done several hours or more after shock, rather than immediately afterward. An example of work in the immediate postshock period, however, is that of Lunn and Trolle (4), who studied mental abilities 10, 30, and 60 minutes after shock. They note very poor retention for materials presented shortly after shock. Thirty minutes afterward, however, they report that 85 per cent of their patients were able to give "a satisfactory description of the content of a picture and at the same time express a comprehension of the point of the picture" (4, p. 376).

The Problem

Up to this time, relatively little research attention has been given to one of the crucial factors in the ability to profit from psychotherapy—namely, *learning*. This is especially true of learning in the immediate postshock period, when psychotherapy is often instituted. As a partial answer to this lack of information, the present research was designed to study, under controlled conditions, the pattern for recovery of learning ability after shock.

One possible, and no doubt partial, interpretation of the therapeutic effect of convulsive shock is that it temporarily disrupts or disorganizes abnormal mental patterns and permits the substitution of different patterns of a more normal or healthy character. We are well aware of the fact that a startling event can shock a normal person "out of his complacency" and cause him to see familiar things in new ways. It may be that convulsive shock does somewhat the same thing for the abnormal person, and jolts him out of his "psychotic rut." If this line of thought be followed, then it would appear that the optimum time to substitute new patterns of thought would be at the point where disorganization is great enough to permit their introduction and yet where learning ability is sufficiently recovered to permit their retention. In other words, if the above interpretation has any validity, the best time for a reeducative therapy would be as soon after shock as a reasonable degree of learning ability has returned.

Subjects. Subjects in this investigation were five female and three male patients at Madi-

---

[1] The cooperation of Ott B. McAtee, M.D., Superintendent, William E. Murray, M.D., Asst. Superintendent, and the staff of Madison State Hospital in making their facilities available for this study is hereby gratefully acknowledged.

son State Hospital who were receiving electric convulsive therapy. They were all literate, in relatively good contact, and cooperative. Their ages ranged from 27 to 62, the average being 37.1. Six were schizophrenic, one manic-depressive, and one involutional.

Experimental materials. Three types of learning materials were developed and used —pictorial, verbal, and numerical. To test visual object learning, a series of 16 line drawings was prepared. Each drawing contained a mannikin surrounded by certain objects. Fifteen details were catalogued in each picture, such as type of hat, tie, etc. on the mannikin, and house, balloon, and other objects surrounding the mannikin. Five different objects or conditions were systematically varied at each of the 15 locations on the different pictures. For instance, the objects on the right side of the mannikin were a bicycle, a clock, a suitcase, a tree, and a truck; and the mouth of the mannikin was shown variously as open, straight, corners up, corners down, and with one corner up and the other down. The subject was shown a picture and each detail was pointed out and verbally called to his attention with the request that he remember as many as possible. Each picture also contained a different mental hygiene aphorism having eight words and five ideas, which was attributed to the mannikin. An example was, "Interests outside make a person far happier inside." Patient motivation for the task was good and several seemed to enjoy it. Indeed, success with this procedure would seem to dictate its wider use as a measure of learning ability.

To test verbal learning, 16 lists of paired adjectives were developed. Each list contained 11 pairs. They were obtained from a list of 300 pairs developed by Melton and Safier (2), and were relatively equivalent in terms of the similarity ratings given them by 96 university students. A list was slowly read over once to a subject with the instruction that he should try to remember as many of the pairs as possible. Thereupon he was handed a typewritten list containing the first adjective in each pair ranged on the left margin, and on its right the second member imbedded among three other synonymous adjectives. For instance, the pair *Pleasing-Charming*, appeared on the typed list given to him as follows: *Pleasing: Pleasant, Gracious, Charming, Winning*. The subject was asked, "Which word did I read with *Pleasing?*" The same procedure was employed with each of the other pairs.

To investigate numerical learning, a series of 16 lists of numbers was prepared. Each list contained two samples each of two-digit through eight-digit numbers, arranged as in the Wechsler-Bellevue Intelligence Scale. Starting with the shorter numbers, subjects were asked to repeat each one immediately after it was read to him. His score was the highest number he could accurately repeat. No attempt was made to go further when both samples of a given length were failed.

Procedure. Subjects were enlisted by explaining to them that they had been chosen for a series of memory tests. One practice trial was given on each of the three types of materials to familiarize the subject with the procedure. The order of presentation was picture, paired adjectives, and numbers.

A control series of tests was run with each subject one to three days before shock was begun. Runs of the three types of tests were given six different times. Some hours to a day intervened between the first three and the last three runs. A short break of a minute or so was introduced between runs.

Experimental testing was begun as soon after shock as the patient was able to get up and go to a nearby examining room. Two runs of tests were given in succession. A third run was given after the patient had had lunch. This procedure was followed after three shocks—in practically all cases the first three. This made a total of 9 experimental runs as compared to 6 control runs. Runs were rotated so that four subjects got the test series in the order 1 to 15, and the remaining four in the reverse order. The time for beginning of experimental testing after shock varied from 10 to 30 minutes, with an average for all subjects of 18½ minutes. Several patients became nauseated, showed confusional retardation, or displayed paranoid resistance. All these caused testing delay.

Results

Table 1 shows clearly the effect of shock on learning performance. A profound interference with learning ability is shown in the first test run after shock, which was an average of 18½ minutes later. The critical ratios of differences between *experimental* and *total control* means are highly significant except in

Table 1

Scores on Control and Experimental Series

	Control runs				Experimental runs		
Test and measure	1	2	3	Total	1	2	3
15 Mannikin Items							
Mean	10.8	11.3	11.4	11.2	2.5	8.3	12.2
Percentage of success	72	75	76	75	17	55	81
SD				2.21	2.96	2.87	2.21
CR with control total					12.39	4.23	1.82
5-Point Mannikin Aphorisms							
Mean	3.6	2.7	3.2	3.2	.4	1.8	3.4
Percentage of success	72	54	64	64	8	36	68
SD				1.87	1.26	1.87	1.73
CR with control total					7.26	2.89	—
11 Paired Adjectives							
Mean	7.6	7.3	8.6	7.8	4.5	5.5	7.7
Percentage of success	69	66	78	71	41	50	70
SD				2.47	1.66	2.22	2.03
CR with control total					6.76	3.99	—
8 Numbers							
Mean	6.1	6.1	6.0	6.0	5.9	6.0	6.5
Percentage of success	76	76	75	75	74	75	81
SD				.93	1.38	1.23	1.04
CR with control total					—	—	1.26

the case of numerical learning. Interference is noted in less but still significant degree on the second experimental run, except in the case of numerical learning. The second testing run began an average of $38\frac{2}{3}$ minutes later. Learning performance has returned to preshock levels by the third testing run, which was an average of $2\frac{2}{3}$ hours later.

Discussion of Results

It is clear from these results that electric shock profoundly interferes with learning abilities. Greater interference is shown for abstract than for more concrete materials. The most extensive interference occurred in the case of the aphorisms, which were the most abstract materials used in this experiment. During the first test run after shock, only two persons of the eight could remember any part of the aphorisms. One of these two could not be tested until 30 minutes after shock. The least interference occurred in the case of numerical learning and report. This task represented the lowest degree of abstraction. Hetherington (1) also found little influence of shock on immediate number recall.

Learning was measured in this experiment by means of immediate report. No attempt was made to determine whether materials were retained over a longer period of time. In a re-educative therapy program, long-term retention would be necessary, of course.

So far as learning ability is concerned, it would appear highly unprofitable to attempt a re-educative type of therapy immediately after shock. This is deduced from individual as well as group scores. While individual variations in recovery rate were smoothed by combining them to produce a group picture, all individual patterns of recovery closely paralleled the group pattern.

By two to three hours after shock, results show that learning ability has returned to preshock levels. It is quite probable that control level ability may return even earlier. Timing for the third test run was set at the completion of the lunch hour for patients and so this was not determined. The shortest period for beginning the testing of any patient on the third run of tests was 1 hour and

55 minutes after shock. In this one case learning ability was back to preshock level.

Insofar as present results are concerned, re-educative therapy is contraindicated for most patients during the first hour after shock. Recovery rates varied for different patients, but preshock learning performance is recovered somewhere between 1 hour and $2\frac{1}{2}$ hours after shock. According to the theory enunciated earlier in this article, somewhere within this interval would seem to be the optimum time for the institution of re-educative efforts.

Summary

Using pictorial, verbal, and numerical materials, the learning ability of eight mental patients was tested both before and at stated intervals immediately after administration of electroconvulsive shock. Results showed a profound interference of shock with learning ability. Interference was greater for abstract than for more concrete learning materials. Learning abilities were recovered in a gradual fashion, with preshock level being restored not later than two to three hours after shock. Implications of the results for a re-educative type of therapy are noted.

Received January 3, 1956.

References

1. Hetherington, R. Psychological changes due to E.C.T. as measured by a battery of tests given repeatedly. *J. ment. Sci.*, 1952, **98**, 334-341.
2. Hilgard, E. R. Methods and procedures in the study of learning. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 517–567.
3. Kalinowsky, L. B., & Hoch, P. H. *Shock treatments, psychosurgery*. (2nd Ed.) New York: Grune & Stratton, 1952.
4. Lewis, N. D. C. Psychiatry. In R. P. Mackay, N. D. C. Lewis, & P. Bailey (Eds.), *The 1949 year book of neurology, psychiatry, and neurosurgery*. Chicago: Year Book Publishers, 1950. Pp. 209–423.
5. Stone, C. P. Losses and gains in cognitive functions as related to electro-convulsive shocks. *J. abnorm. soc. Psychol.*, 1947, **42**, 206–214.
6. Zubin, J. Objective studies of disordered persons. In T. G. Andrews (Ed.), *Methods of psychology*. New York: Wiley, 1948.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

Note on "Changes in Psychoneurotic Patients With and Without Psychotherapy"

Desmond S. Cartwright
University of Chicago

A recent paper in this Journal by Barron and Leary (2) reports a comparison of changes in a treatment group of psychoneurotic patients with changes in a nontreatment group. Groups were matched for type and severity of disorder, prognosis, socioeconomic variables, and interval of time between pretests and posttests.

Of the patients in treatment, 42 received individual and 85 received group psychotherapy. The measures employed were MMPI scales, including a prognostic scale (*Es*) developed by Barron (1). The crucial data showed mean differences in scale scores between pretests and posttests. Comparison of these data by critical ratio analysis led to the conclusion that: "The therapy patients did not improve significantly more than did the waiting-list controls" (2, p. 245).

This conclusion answered the question: ". . . is there significantly more change with regular and formally defined psychotherapy than without it?" (2, p. 245). Statistically, the answer given states that the treatment and nontreatment samples must be considered to be drawn from the same parent population in regard to mean changes. The presence or absence of psychotherapy makes no difference.

However, from the data presented (2, p. 243), it appears that the samples differ as to the dispersion of changes (*SD*s). Table 1 presents the relevant data for the individual psychotherapy group and the nontreatment group, together with *F* ratios computed from the variances.

Table 1 suggests with high probability that the two samples are not drawn from the same parent population in regard to the variance of changes on scales *F*, *Hs*, *D*, *Pt*, *Sc*, and *Es*. This means that the presence or absence of individual psychotherapy makes a considerable difference.

The implication of the results in Table 1 is that in fact there was "significantly more change with regular and formally defined psychotherapy than without it."

Table 1

Variances of Discrepancy Scores on MMPI Scales for Individual Psychotherapy and Nontreatment Groups

MMPI Scale	Individual psychotherapy ($N=42$) V*	Nontreatment group ($N=23$) V	*F*	*p*
1. Lie	19.89	23.43	1.18	—
2. *F*	215.21	22.94	9.38	.01
3. *K*	55.95	31.70	1.76	—
4. *Hs*	127.46	64.16	1.99	.05
5. *D*	244.30	93.32	2.62	.01
6. *Hy*	113.21	87.80	1.29	—
7. *Pd*	155.00	89.68	1.73	—
8. *Pa*	111.94	68.06	1.64	—
9. *Pt*	208.51	73.27	2.85	.01
10. *Sc*	272.91	74.13	3.68	.01
11. *Ma*	126.79	75.34	1.68	—
12. *Es*	43.56	14.82	2.94	.01

* Variances computed from *SD* data reported by Barron and Leary (2, p. 243).

The greater amount of change, however, was not all in one direction. Thus the authors' conclusion that "The therapy patients did not improve significantly more than did the waiting-list controls" is correct but insufficient. For many scales the variance results suggest that mean differences between the groups are absent because differences of two kinds, opposed in sign, are present. It seems that some therapy patients *deteriorated* to a greater extent than did the waiting-list controls, while some therapy patients *did improve* significantly more than the controls.

Received November 9, 1955.

References

1. Barron, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.,* 1953, **17**, 327–333.
2. Barron, F., & Leary, T. F. Changes in psychoneurotic patients with and without psychotherapy. *J. consult. Psychol.,* 1955, **19**, 239–245.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

A Short Form of the Wechsler-Bellevue and the Army Classification Battery as Measures of Intelligence[1]

Donald B. Murphy
Army Medical Service School, Ft. Sam Houston, Texas

and Robert D. Langston
Br. U. S. Disciplinary Barracks, Ft. Crowder, Mo.

Although the Army Classification Battery and short forms of the Wechsler-Bellevue have been used as measures of intelligence, a review of the literature indicated the application of previously published short forms to an Army disciplinary barracks population was somewhat questionable. The present study was designed to investigate the validity of a short form of the W-B and of Aptitude Area I (AAI) of the Army Classification Battery as measures of intelligence for this population.

For the present study, the records of 515 consecutive releasees from the Br. U. S. Disciplinary Barracks, Ft. Crowder, Missouri, were reviewed. The ages of this group of male prisoners ranged from 16 to 41 with the mode falling at the 20–24 age grouping. Eighty-five per cent worked at unskilled labor or had no occupation prior to induction. Negroes comprised 30% of the group.

The W-B full scale IQ, short form IQ based on the Information, Comprehension, and Block Design subtests, and the scores on the AAI, which has a normative population mean of 100 and *SD* of 20, were obtained for each case.

The mean of the full scale W-B was 93.2 with an *SD* of 16.1; the mean of the short form was 92.8 with an *SD* of 16.9; and the mean of the AAI was 81.7 with an *SD* of 22.2. The short form of the W-B and the AAI correlated .91 and .83, respectively, with the full scale W-B.

The advantage of the short form, since it is a reliable predictor of full scale scores, is that testing time is reduced, permitting rapid processing and more time for evaluation of nonintellectual factors.

The *S*s used in this study had a significantly lower IQ ($p < .01$) than Wechsler's normative group and differed from Wechsler's population with respect to occupational status, education, and race. It is interesting to note that the short form used in this study has been reported to have the highest correlation with the full scale W-B for Wechsler's normative group, thus indicating the stability of these subtests despite differences in socioeconomic, educational, and racial backgrounds.

The mean for the AAI was significantly lower ($p < .01$) than the AAI mean (90.7) estimated on the basis of the *S*s' W-B scores. Marked increase in the variability of AAI scores below 75 suggested that many of the low scores resulted from poor motivation at the time the test was administered rather than from low intelligence. When AAI scores below 75 were discarded, the mean of the AAI scores (93.4) closely approximated the estimated AAI mean and correlated .88 with the full scale W-B.

Brief Report.
Received April 3, 1956.

---

[1] An extended report of this study may be obtained without charge from Capt. Donald B. Murphy, AMSS, Ft. Sam Houston, Tex., or for a fee from the American Documentation Institute. To obtain it from the latter source, order Document No. 4970 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

"Reliability of the Blacky Test": A Reply to Charen

Gerald S. Blum
University of Michigan

After reading Charen's brief report on the retesting of 25 tuberculous patients, printed in the February issue of this Journal (5), I was left with two major misgivings: (*a*) the title is a misnomer inasmuch as the technique was used improperly; and (*b*) the author's reasoning is faulty. I shall elaborate these points below:

1. The Blacky Pictures are designed to yield conflict scores for a series of psychosexual dimensions. As stated in the Manual (3), four separate sources contribute to each dimensional score: spontaneous story, inquiry items, related comments, and cartoon preferences. The underlying conception of reliability, upon which the instrument was constructed, assumes that these sources tap the conflict area at varying levels and in combination provide the best possible estimate of conflict intensity. The revised scoring system (4) referred to by Charen makes clear this necessity for including all four sources in deriving a reliable dimensional score. Instead, Charen arbitrarily singled out for investigation one source, multiple-choice items from the inquiry, which is stated to be closest to consciousness and, therefore, most easily distorted by the subject. Also the technique was administered incorrectly, since spontaneous stories were not obtained prior to the responses to multiple-choice items.

2. From the facts that the Rorschach and paper-and-pencil tests indicated no "basic changes" in personality whereas the Blacky items did show some (basic?) changes, Charen infers only two possible alternatives: "Either the Blacky questions tapped personality levels not reached by the other tests or did not have good reliability." He rejects the first alternative by a statement of faith in the Rorschach. I prefer to call upon evidence contained in a study by Berlow (2), who compared Blacky scores (in a sample of patients whose responses had previously been meaningfully related to theories of paranoid schizophrenia [1]) to psychosexual indicators on the Rorschach. In addition to using judges to assess psychosexuality from the Rorschach, Berlow conducted a systematic and exhaustive analysis of signs, patterns, and content. No significant relationships to the Blacky were established. Had Charen employed the Blacky correctly, there would still be no reason to expect congruence with Rorschach findings. Berlow's results are not introduced to set off a debate on the relative merits of the two instruments, but simply to point out that there is no place for the Rorschach in Blacky's orbit of construct validation.

Finally it is important to keep in mind that even when the technique is administered and utilized correctly, the retest method, reflecting as it does an unknown mixture of changes in set, familiarity with items, and perhaps personality changes as well, cannot properly be employed to study reliability of responses to the Blacky Pictures.

Received March 23, 1956.
Early Publication by the Editor.

References

1. Aronson, M. A study of the Freudian theory of paranoia by means of the Blacky Pictures. *J. proj. Tech.*, 1953, **17**, 3–19.
2. Berlow, N. Psychosexual indicators on the Rorschach test. Unpublished doctor's dissertation, Univer. of Michigan, 1952.
3. Blum, G. *The Blacky Pictures*. New York: Psychological Corp., 1950.
4. Blum, G. Revised scoring system for research use of the Blacky Pictures. Ann Arbor, Michigan: Author, 1951 (mimeographed).
5. Charen, S. Reliability of the Blacky Test. *J. consult. Psychol.*, 1956, **20**, 16.

Journal of Consulting Psychology
Vol. 20, No. 5, 1956

A Reply to Blum

Sol Charen
Montgomery County Mental Health Clinic, Rockville, Maryland

Blum pinpoints emphasis on my title to suggest that the technique I used is incorrect. A title is only intended to furnish brief information about an article. Mine could have been clearer. It is also necessary to read the article to find out what it is about and not rely solely on the title. Clearly stated throughout my brief report, "The Reliability of the Blacky Test," is the fact that it deals with the low reliability of the multiple-choice questions of the Blacky test, as found in my doctoral research (2). This is all my article deals with, not the total Blacky test, and Blum is unwittingly in accord with my text when he states in his reply ". . . multiple-choice items from the inquiry, . . . [are] closest to consciousness and, therefore, most easily distorted by the subject." As he points out, though, the multiple-choice questions do furnish one of the four sources which contribute to each dimensional score. The question of reliability of the total Blacky test is therefore involved when a major score source may be questionable.

Since spontaneous stories were not obtained from my chronically ill patient subjects I did not follow his technique. How this omission affected their responses to the multiple questions for each cartoon I do not know. But Blum is on shaky ground when he implies that this omission somehow affected the reliability since he has no reliability studies of his own for either the total Blacky test or any of its four scoring sources. Without such data, his assumption is purely hypothetical that the omission of the spontaneous stories somehow affected the reliability in the test-retest situation.

Blum has lifted a sentence out of context. Nowhere did I use the Rorschach test to validate either the Blacky test or its multiple-choice items. The Rorschach and fifteen paper-and-pencil tests were used to evaluate whether my patients in a four-month period showed test evidence of basic personality changes as a result of chronic illness and extensive hospitalization. No test evidence of such changes was found. Only the multiple-choice items of the Blacky test indicated that some changes took place in my experimental subjects. These, in general, suggested they were nonregressed during illness testing and were regressed when retested shortly before discharge from the hospital. I believe I am consequently justified in assuming that the equivocal results found in the Blacky multiple-choice items may have been due to the factor of the low reliability which was found to exist.

One final comment about the last paragraph of Blum's reply is in order. The question of set should not be involved in the Blacky test as it might be in other personality tests since his scoring allows for evasiveness or attempts to use defenses. Familiarity with items is not likely to be a problem when a four-month or longer period elapses between the test and retest. Finally, since the Blacky test purports to measure, "the deeper recesses of personality" (1), namely the genetic psychosexual dimensions which should not change in the individual unless psychoanalysis or a marked life experience occurs, personality change is also not a factor which should prevent use of the retest method.

Received June 25, 1956.
Early publication by the editor.

References

1. Blum, G. S., & Hunt, H. F. The validity of the Blacky pictures. *Psychol. Bull.*, 1952, **49**, 238–250.
2. Charen, S. Regressive behavior changes in the tuberculous patient. *J. Psychol.*, 1956, **41**, 273–289.

PSYCHOLOGICAL TEST REVIEWS

Doll, Edgar A. *Vineland Social Maturity Scale.* Minneapolis: Educational Test Bureau, 1947.

Reviewed by

Morris Krugman

Board of Education, New York, N. Y.

An instrument that has been in use for thirty years is hardly news. The first article about the *Vineland Social Maturity Scale* by Edgar Doll appeared in 1935, after almost ten years of use and development at the Vineland Training School; the last revision was made in 1936, and the manual accompanying the current form of the scale was published in 1947. Even a casual inspection of psychological literature reveals more than a hundred titles discussing or mentioning use of the scale, and many more are undoubtedly available.

Something new, however, has been added, with the publication of *The Measurement of Social Competence* by Edgar A. Doll in 1953. For the first time in thirty years, the rationale, the details of the standardization process, scoring samples, methods of interpretation, and other pertinent data usually supplied when psychological instruments are first presented for general use, are made available for the *Vineland Social Maturity Scale*. Although the brief 1947 manual mentions the care with which the standardization process was carried out, any real idea of the meticulousness of this process can be gained only from the voluminous 1953 manual. Each of the 117 items of the scale is described in minute detail with possible variations, normative curves are presented for each item, and means, medians, *SD*s and *CR*s supplied. Line drawings illustrating typical behavior are added. One wonders whether the treatment of data was overelaborated, since the norms are based on ten males and ten females at each age level from 0–1 to 30, a total of 620 subjects.

The scale is too well known to require detailed description here. Briefly, it consists of 117 items, scaled in order of difficulty and arranged by years from 0 to 30. The items are distributed among six categories: self-help (general, eating, and dressing), self-direction, occupation, communication, locomotion, and socialization. "Each item is conceived as representing a general growth in social responsibility which is expressed in some detailed performance as an overt expression of that responsibility. Consequently, the value of the detailed items is to be determined principally by the extent to which they reflect this personal independence in personal activities, in respect to which the detailed performances are otherwise relatively unimportant." The items "reflect progressive freedom from need of assistance, direction, or supervision on the part of others."

Although direct rating by the examiner is not ruled out if opportunity for observing or knowing the subject is available, the usual method of use is by interviewing an informant who is thoroughly familiar with the behavior of the subject. The fact that the subject being examined and rated need not be present or observed is both an advantage and a disadvantage. The obvious advantage is the case of applying the scale by questioning a generally mature, competent adult about an immature, incompetent, or handicapped individual. The disadvantage lies in the possibility of gross distortion and incorrect rating. The reviewer has always experienced a feeling of greater confidence about ratings when the subject was seen even briefly after the usual rating by means of an informer, and when a few items of the scale have been spot checked.

Usually, the publication of basic data about a clinical instrument serves to clarify issues and strengthen the use of the instrument. This is not the case with the *Vineland Social Maturity Scale*. That the scale is a very useful clinical adjunct with many different types of subjects is attested to by the almost uniformly favorable literature over the past twenty years. This reviewer has used it for many years and has tremendous respect for it. But this respect is not increased by careful reading of the new manual. The small number of cases at each level has been mentioned and indeed was known from the earlier manual, but the elaborate statistical and graphic treatment of each item is not convincing, since, in spite of the fact that continued appraisal of each item was made statistically, experimentally, and clinically, not one of the 117 items has been shifted up or down, nor have items been replaced or modified. Equally questionable is the upper end of the scale, intended for normal and superior adolescents and adults. Although based on normative data, this part of the scale could well be an extrapolation of the lower end of the scale, and it hardly makes sense. An ingenious method of measuring defects in social maturation of children and adults, and indicating the level of social development of children who are not necessarily retarded in development, the Vineland Scale is artificial when applied to normal or superior adolescents or adults. Most users of the scale simply ignore the upper levels.

The *Vineland Social Maturity Scale* has its greatest usefulness in distinguishing between real and pseudo-feeblemindedness, in providing a measure of social development in studies of child growth and development, and in appraising levels of social function among handicapped individuals of many kinds and in many types of pathology, particularly in cases of arrested development or deterioration. It has made a permanent place for itself in clinic and laboratory, and newer approaches to the appraisal of social maturity have not succeeded in replacing it. The fact that it has not been found necessary to modify or "improve" the scale over these many years can be accepted as evidence of the carefulness with which it was originally constructed, and the keen clinical insight that obviously determined the original construction and standardization.

Doll, Edgar A. *Vineland Social Maturity Scale*. Minneapolis: Educational Test Bureau, 1947.

Reviewed by

Edwin S. Shneidman

VA Neuropsychiatric Hospital, Los Angeles

The *Vineland Social Maturity Scale* is not a psychological test in the usual sense, nor does this review pretend to be the usual type of test review. Most tests are samples of modes of reacting, perceiving, or adjusting, or are extrapolations from segments of behavior. The *Vineland Social Maturity Scale*, on the other hand, attempts to measure behavior directly. Although Doll says that *"The scale is not a rating scale and scores are not to be based on mere opinions"* (italics his), nevertheless the scale would seem to be in the same tradition and of the same genre as the Gesell *Development Schedules*, the Malamud and Sands *Psychological Rating Scale*, the Ferguson-McReynolds *Hospital Rating Scale*, etc. The *Vineland Social Maturity Scale* is a well known (to more than one generation of clinicians) measure of the social and adaptational skills, habits, traits, coping behaviors, and interpersonal reactions of subjects between the ages of 0–1 and 25 plus. Doll indicates that the purpose of his scale is "To provide an over-all measure of personal adequacy in terms of social maturity." He further indicates that the scale is a "descriptive analysis of maturation in social competence, as a standardized schedule of behavior development in terms of its habitual expression." He feels that it has use in establishing a standard schedule of normal development, in measuring deviations which may be significant in such problems as delinquency, amentia, adoption of children, etc., in measuring improvement following training or therapy, etc.

There are 117 items in the scale, arranged in order of average age norms. The items have been chosen to report some aspect of the individual's ability to look after his own practical needs. Social quotients (SQ) "generally comparable to Binet MA and IQ can be obtained from the sum of the items scores." The items themselves range from such as "'Crows'; laughs" (0–1) to "Advances general welfare" (XXV). The items at the upper end of the scale (ages 20 up) such as "Uses money providently," "Contributes to social welfare," "Systematizes his own work," "Inspires confidence," "Promotes civic progress," "Advances the general welfare," etc. seem extremely hard to evaluate. Nor does Doll's injunction that *"The informant does not make the scoring judgment. This is done by the examiner after obtaining from the informant as much detail as practicable regarding the behvaioristic facts which reveal the manner and the extent of the subject's actual performance on each item"* (italics his) really circumvent the essential problem of observing and rating items that are essentially difficult to rate.

The main purpose of this summary is neither to describe the *Vineland Social Maturity Scale* nor to present a survey of the published research on it; it is rather to discuss some aspects of its implied rationale, and some cautions in its application. Certainly the careful use of the scale to distinguish those subjects who do deviate from the norms of "social maturity," so that further study of those individuals can be made, would be a useful and rewarding application of this instrument. However—and this is the burden of this review—the *Vineland Social Maturity Scale* lends itself easily to other interpretations.

Ordinarily, when one uses an intelligence test in order to obtain an IQ, the main purpose of such an effort is to ascertain what the IQ is, although the implicit assumption may be that the higher the IQ, the better. This implicit assumption operates in a much more subtle and in a more questionable way in the case of the *Vineland Social Maturity Scale*. Specifically, the assumption is that social behavior (including social motivation) is desirable as it approaches social conformity. Further, the implicit assumption is that optimal behavior lies in adjustment to a standardized society and conformity to the "normal." Without becoming embroiled in the intricacies of the late Robert Lindner's concepts of rebellion and salutary nonconformity, it still remains that the *Vineland Social Maturity Scale* deemphasizes the individual's fealty to his own image, his strength to be himself, his positive (that is creative or effective) idiosyncracies, his areas of adaptation which may bring him pleasure, or the application of his resources (within his total personality organization) toward creativity or self-actualization. The recent work of Jean MacFarlane at Berkeley and Lois Murphy at Sarah Lawrence would seem to indicate that the conforming person is not necessarily the best adjusted person. One is reminded of a section of D. B. Klein's book *Mental Hygiene* entitled, "In Defense of the Shy Child" (who does not necessarily become schizophrenic) and one wonders if a section should be added in defense of the nonconforming child (who need not necessarily become maladjusted, psychopathic, or feebleminded).

The point of all the above statements is that the *uncritical* use of the *Vineland Social Maturity Scale* may imply an unwarranted assumption of the merits of conformity and standardization and to permit a confusion between aberrant or pathological nonconformity (schizophrenia, for example, at one extreme). To make these assumptions would be to gainsay the infinite variety and the rich diversification of talents and gifts of human subjects.

As a clinical psychologist—and as a father of four—I am not sure that I know what the "best" age for toilet training, for example, really *is*. I would also ask whether the *child* has a problem if he isn't toilet trained by "that" age or rather cannot the parent create problems by imparting parental and societal anxieties and concerns—*his* problems—to the child.

The major reference for the *Vineland Social Maturity Scale* is Doll's *The Measurement of Social Competence* (Minneapolis: Educational Test Bureau, 1953). Two shorter references are the *Manual of Instructions* (published by the Educational Test Bureau, dated 1947) which cites 59 references, and Dr. Doll's chapter (in the second part of Chapter 20) on his scale in Weider's *Contributions Toward Medical Psychology* (New York: Ronald Press, 1953).

Rotter, Julian B., & Rafferty, Janet E. *The Rotter Incomplete Sentences Blank (College Form)*. New York: Psychological Corporation, 1950.

Reviewed by

Louis Long

The City College of New York

The Rotter Incomplete Sentences Blank (ISB) "is a semistructured projective technique in which the subject is asked to finish a sentence for which the first word or words are supplied" (Manual, p. 3). Historically, the incomplete sentence method is closely related to word association tests, and the ISB has a long history of evolution. Sentence forms used by Rotter and Willerman, Shor, Hunt, Holzberg, Rohde and Hildreth, and Payne have had an effect upon the ISB. As in all projective situations it is assumed that the subject in completing the sentences will reflect something of his own personality—his fears, desires, wishes, anxieties.

The ISB consists of 40 "stems" that are to be completed by the subject. The completions are then scored by comparing them with responses that have been assigned a scale value varying from 0 to 6. The total score is an index of maladjustment.

The ISB may be used in two very different ways: (*a*) as a gross screening instrument; and (*b*), as a clinical tool. It seems desirable to evaluate these two uses separately.

Evaluation of the ISB as a screening instrument. Biserial coefficients of 0.50 and 0.62 between scores and classifications of adjusted and maladjusted students are presented in the manual to show how effective the test is when used as a screening instrument. The small number of cases upon which these data are based should, however, warn clinicians to proceed with caution in using the ISB for this purpose. Each person using the test will do well to study the effectiveness of it in his own local situation. This in essence means that the ISB must be validated locally which will necessarily require setting up criterion groups of adjusted and maladjusted individuals. In their manual, Rotter and Rafferty describe in detail the method they used to establish their criterion groups. It is easy to sense their dissatisfaction with the use of ratings of adjustment or maladjustment as supplied by instructors teaching classes in effective study and in mental hygiene, or by advanced student clinicians. In the reviewer's opinion, the difficulty of establishing satisfactory criterion groups will inevitably make it impractical for many colleges to use the ISB. If, however, the aggressive research program which seems to be underway at present is continued it may eventually be possible for colleges to use the ISB as an aid in screening students for possible referral to a counseling agency.

Evaluation of ISB as a clinical tool. To facilitate the use of the ISB as a clinical instrument the authors of the Manual have explained their scoring procedure in detail, have given many examples of the various kinds of responses, have presented practice cases along with the correct scoring of the responses, and have given a clinical interpretation of six records. Each of the six cases begins with a brief background sketch of the individual, followed by a record of his responses on the ISB, and an analysis of the case in terms of familial attitudes, social and sexual attitudes, general attitudes, and character traits. The reports conclude with a summary, and a discussion of the contribution to the ISB. These cases are presented to illustrate the potential value of the ISB and it is emphasized by the authors that only *some* of the possible interpretations are included in the analysis. As one reads the analyses and studies the responses other interpretations often seem plausible. This in a sense is an indication of the value of the test since one of the main purposes of a projective situation is to furnish leads that may be explored in an effort to gain clinical insights. The authors are effective in showing how the ISB adds to their understanding of clients. It would be interesting to know how the ISB's contribution compares with that of the Rorschach or the TAT. Specific information about the relative contribution of the several projective devices is urgently needed.

The authors of the manual are careful to point out the limitations of the ISB and to remind readers that all data presented relate to the college form. It is somewhat unfortunate that high school and adult forms are listed as being available. Until such time as manuals and norms are available it would seem desirable to keep such listings out of test catalogues.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Second-Order Personality Factors in the Questionnaire Realm

Raymond B. Cattell
Laboratory of Personality Assessment and Group Behavior, University of Illinois

Although there has been much emphasis on the potential importance of broader, second-order personality factors, by Thurstone and others, and even some claims to recognizing particular second-order factors by Eysenck (20) and the present writer (2, 4), there has not existed until this year any very firm basis of evidence for conclusions. The present writer's earlier attempt (4) was admittedly premature, while Eysenck's extensive arguments for a second-order neuroticism factor and "extraversion" factor (20) seem partly based on the misapprehension, mentioned below, that second-order factors can be reliably obtained from factoring correlations among the actual scores on the existing measures for primary factors. There are, indeed, three rather widespread misapprehensions, dealt with below, which may confuse research on second-order factors.

A second-order factor is a factor among oblique primary factors. That is to say, it is obtained by factoring a matrix of correlations among pure measures of primary factors and rotating to simple structure. The second-order factors will be less numerous than first orders, or "primaries," and will tend to cover wider areas of behavior. For example, the general ability factor, on which the IQ is based, covers more than the individual primary abilities of Thurstone, from which it is derived as a second-order factor (30).

The theoretical and practical value of second-order factors, relative to first-order factors, is an issue with partisans on both sides. In the theoretical realm some, like Hsü (25) see second-order factors as "super factors," more permanent and more deeply rooted in physiological or sociological fact than first-orders. Others (6, 24, 33) consider the primaries as the psychologically-real concepts and second-orders only as more transient organizing principles among them. The former is illustrated by the theory that *g* is a function of the total effective number of cortical neurones, but that the primary abilities merely represent particular training investments of this general relation-perceiving capacity, developed in particular sensory or symbolic areas. The latter, i.e., the alternative interpretation, is illustrated by the present writer's earlier explanation that the observed organization of certain personality factors (2) represents only the population correlation produced among them by selection according to the social status conditions of the particular country concerned. (See also Fisher, 21.)

Regardless of theory, the practical psychometrist can predict test performance either from first- or second-order factors. But here, too, there are preferences, as shown, for example, by the British predilection for working with Spearman's *g*—the second-order factor in abilities (33)—and the American preference for primary abilities. If we rule out specifics as too numerous to be useful, prediction from primaries accounts for the greatest possible actual test variance, but from second-order factors there is a wider field of prediction, in terms of one factor (6, 30).

Second-Order Factoring

For the psychologist, as contrasted with the statistician, the importance of factors lies beyond immediate prediction. It concerns isolation of psychological concepts and the development of psychological laws and insights, which extend the understanding of behavior beyond the immediate, static, measurement situation. Here, both the first- and the second-order structure are parts of the total necessary knowledge, as will be exemplified in the present study. However, as indicated in the introductory remarks, the second-order factors must for this purpose meet certain conditions not sufficiently recognized in some current studies.

In the first place, it does not suffice to take any existing test or battery measuring the factors, to intercorrelate the factor scores, and analyze. Today, and for many years to come, the individual factor measures will be substantially impure. Each will contain, in addition to the inevitable specific, some contamination with other factors present in the battery. The primary measures may, also, include in common some factors, e.g., response sets, which have nothing to do with the intended primary measures per se. Consequently the factoring of these actual primary factor test scores, presuming that each is split in two, or increased in number in some way to give a sufficiency of variables for determining the number of common factors involved, will yield the primaries again, together, perhaps, with the spurious common factors among them due to dimensions extraneous to the primary dimensions themselves, e.g., to the response format of the battery.

This theoretical expectation has actually been verified in Thurstone's (31) and Lovell's (27) work on Guilford's factors (24), where, if the battery for each factor is split, and each factor reliability is put in the matrix, the analysis is shown to yield the primaries again with tolerable precision. This result will be readily appreciated by anyone who has seen the complete discrepancies between the correlation between the raw scores of any two factor measures and the correlation found from measuring the cosine of the angle between the factors on their simple-structure plot. The former, dependent on a cluster of items, is pulled all over the place by the addition of fresh items, with their particular contaminations; but the latter remains firm. For the hyperplane is fixed by variables in *many* dimensions that are *not* loaded in the factor in question. However, it must be granted that where factors are already very purely measured by existing tests, with no systematic bias and with effective suppression of peer factors in each factor score, the factoring of actual factor scores, i.e., of one variable for each factor, may yield a reasonable approximation to second-order factors. This seems to have happened with the 16 P.F. test in the large Air Force study (8).

However, the matrix with which a second-order factoring properly begins is the cosine matrix among the simple-structure primary factors. Few oblique rotations to simple structure undertaken in the usual course of determining the form of the primaries are accurate enough for a second-order factoring. Indeed, as French's (22) and other surveys show, perhaps a half of all published factor analyses falter and fall short of their goal—even the goal of defining the first-order factors—for lack of thorough search for simple structure. Even the last couple of over-all, "polishing" rotations, in the usual series of ten to twenty over-all rotations from the original position, will substantially alter the number and nature of second-order factors that would be obtained.

One needs also to be warned against absent-mindedly factoring—as the writer has done before (4)—the matrix of cosines thus achieved, for these are the correlations among the reference vectors. Such a matrix must be inverted to give correlations among the factors. Thus when, as in the usual notation, λ is the obtained transformation matrix, proceeding from the extracted factor matrix, V_o, to the best simple structure matrix, V_n, the required square matrix for second-order factoring is

$$D(\lambda'\lambda)^{-1}D$$

where D is a diagonal matrix the purpose of which is to normalize by rows and columns the obtained inverse (6). In summary, the three pitfalls to be avoided in second-order factoring are: taking primary correlations from raw scores instead of the simple structure plot; using reference vector correlations instead of factor correlations; and, as dealt with below, leaving second-order factors orthogonal instead of seeking simple structure for them also.

The Questionnaire Foundation

The number of primary personality factors in questionnaire responses is still not fully determined, but the recent studies of Guilford (24) and of the present writer (10, 11, 12) converge in indicating that (among adults) about fifteen will account for most of the variance in highly varied collections of items. In the case of the author's studies and those of his research associates, the close coordination maintained between questionnaire and observer behavior-rating factors brings additional confirmation, in that some fifteen factors have also been found in the latter (3, 12).

The fifteen primary personality factors on which the present study is based have since 1949 been embodied in the Sixteen Personality Factor Questionnaire (15), into which, for general utility, general intelligence has been built as the sixteenth factor. The factor structure of the 432 items (466 less 34 intelligence items) comprising the A, B, and C forms of this test has been verified in three studies, each experimentally, statistically, and rotationally independent (5, 10, 11). A wide breadth of behavior reference in the 16 P.F. items has been guaranteed by the historical survey (3) and by the concept of sampling the personality sphere made in the first analysis (5), as well as by examination of 2,002 further items in the two succeeding studies (10, 11).

Hypotheses as to what the fifteen factors may be which so persistently appear have been put forward (3, 15, 22) and debated. However, it is only recently, with the return of external validation results from the 16 P.F. Test (9, 14, 15, 34), in clinical, educational, and industrial use, that these concepts have been evaluated and sharpened. For example, the examination of the degree of inheritance of these factors made on their equivalent forms in the Junior Personality Quiz (13) with eleven-year-old twins, siblings reared apart, etc. (14), has thrown much light on their basic natures. From these studies, it appears that 2 of the 15 factors (*A* and *H*) are hereditary components of the cyclothyme-schizothyme temperament continuum; that others, e.g., surgency-desurgency (*F*), and sensitivity (*I*), are almost wholly environmentally determined; that the factors *C* and *G* correspond to the clinical concepts of ego strength and superego strength; that *E, L,* and Q_1 correspond to the concepts of dominance, paranoid trend, and radicalism, and that *O* and Q_4 characterize two relatively distinct components in anxiety. More precise discussion can be given to these meanings of primaries where it becomes relevant to the interpretation of the second-order factors among them found here.

Correlations Among Primaries

Determination of the cosines (correlations) among the fifteen primary factors was based on the third (10) of the above studies on the 16 P.F., which had benefited from the preceding studies in acquiring more highly loaded items, and, therefore, a clearer factor structure. The correlation matrix was of order 75 × 75 and consisted of 2,775 product-moment correlations on an *N* of 408 subjects (181 undergraduate men and women students and 227 Air Force young men in training). Although the 16 P.F. test was represented by only 75 variables it will be seen by reference to the studies indicated (10, 11) that these truly represent the factor structure of the whole test. The special procedure of *parcelled factor analysis,* by which the 163 highest loaded items in the test are parcelled into 75 short scales of two and three items each (because of the impossibility of handling a 163 × 163 matrix) has been described elsewhere (11). It gives the test as a whole the structure of the 75 parcels, and evaluates each item in the total 16 P.F. against the factors established in this framework.

For purposes of determining the angles among oblique factors with great exactitude, in order to enter a second-order factor analysis, this design of having each factor loaded on only 5 out of 75 variables is particularly valuable. For the even distribution (each factor represented by the same number of variables) and the great preponderance of variables in every hyperplane (70 out of 75 is possible, and was closely approached in our

Table 1

Cosines Among Reference Vectors
Primary Personality Vectors

	(A)	(C)	(E)	(F)	(G)	(H)	(I)	(L)	(M)	(N)	(O)	(Q₁)	(Q₂)	(Q₃)	(Q₄)
(A)	1.01														
(C)	.13	1.00													
(E)	.07	.06	1.00												
(F)	−.01	.05	.00	.99											
(G)	.00	.04	−.02	.05	1.00										
(H)	−.36	−.12	−.18	−.26	.00	1.00									
(I)	−.29	−.04	−.06	.01	.00	.14	1.01								
(L)	−.01	−.01	−.06	−.11	−.04	.26	−.04	1.00							
(M)	.03	.14	.06	.26	−.04	.05	−.14	−.07	.99						
(N)	.08	.01	−.11	−.12	.13	−.05	.26	−.05	−.02	1.01					
(O)	.02	.19	.00	−.04	.02	.13	−.11	−.08	−.06	.04	1.00				
(Q₁)	−.15	.05	−.31	−.02	.11	−.09	.02	.01	−.19	.00	.01	1.01			
(Q₂)	.07	−.01	.18	.22	.02	.03	−.11	−.01	−.11	.03	.08	−.14	1.01		
(Q₃)	.30	−.19	.12	.22	−.19	−.29	−.02	.11	−.04	−.22	.20	.00	.12	1.00	
(Q₄)	.04	.09	.02	.05	−.05	.01	.15	−.15	−.01	−.07	−.17	.03	.01	.23	.99

study) permits a very sure perception of hyperplanes and their angles. Actually the mean percentage of variables obtained in the average hyperplane (between ± .10 loadings), for all factors, came to 91 per cent, which is larger than in any other published or unpublished study known to the writer. Secondly, the determination of each hyperplane by as many as fourteen factors, i.e., in fourteen-space, precludes any serious distortion of angles from the composition of the particular cluster of variables which might happen in this study to define, i.e., to be highly loaded in, any given factor.

The simple structure was actually approximated with relatively little labor by applying Wrigley's quartimax method (28). This gives the nearest approach by an analytical method, and within a Procrustean orthogonal structure, to the eventual oblique structure. However, since this analytic method stops short of the best simple structure, by tending to leave too much variance in the centroid and through failing to follow the natural obliquities, rotation was finished "by hand." Graphical plots of each factor with every other (105 plots per rotation) were made. After four rotations of all factors from the quartimax po-

Table 2

Correlations Among Factors
Primary Personality Factors

	(A)	(C)	(E)	(F)	(G)	(H)	(I)	(L)	(M)	(N)	(O)	(Q₁)	(Q₂)	(Q₃)	(Q₄)
(A)															
(C)	−.19														
(E)	−.02	−.04													
(F)	.16	−.04	.16												
(G)	−.04	.04	−.05	−.02											
(H)	.15	.32	.21	.28	.01										
(I)	.30	−.15	−.06	−.12	.02	−.24									
(L)	.04	−.26	.00	.03	−.02	−.42	.15								
(M)	−.06	−.20	−.08	−.36	.05	−.28	.23	.15							
(N)	−.22	.17	.16	.15	−.08	.26	−.40	−.11	−.13						
(O)	.13	−.44	−.02	.03	−.07	−.40	.23	.35	.17	−.22					
(Q₁)	−.16	−.03	.31	−.03	−.11	.08	−.05	−.02	.18	.08	−.03				
(Q₂)	−.06	−.06	−.20	−.33	−.01	−.24	.19	.08	.28	−.17	.04	.11			
(Q₃)	−.32	.47	−.03	−.14	.18	.42	−.24	−.39	−.08	.33	−.52	.04	−.11		
(Q₄)	.03	−.31	−.01	.02	−.02	−.28	−.03	.33	.09	−.05	.38	−.04	.03	−.42	

sition, a very clear structure was obtained. A fifth "polishing" rotation was carried out to give the best possible determination of angles. The mean hyperplane content indicated above, 91 per cent, was thus obtained, and the significances of the simple structure determinations, tested by Bargman's formula (1), was found to be beyond $p = .001$ for each of the fifteen factors.

The cosine matrix, $C = \lambda'\lambda$, among the reference vectors at this point is given in Table 1 and its inversion and normalization to the correlations among the factors, $K = DC^{-1}D$, in Table 2. It will be observed that as usual the intercorrelations among the factors are higher than among the vectors.

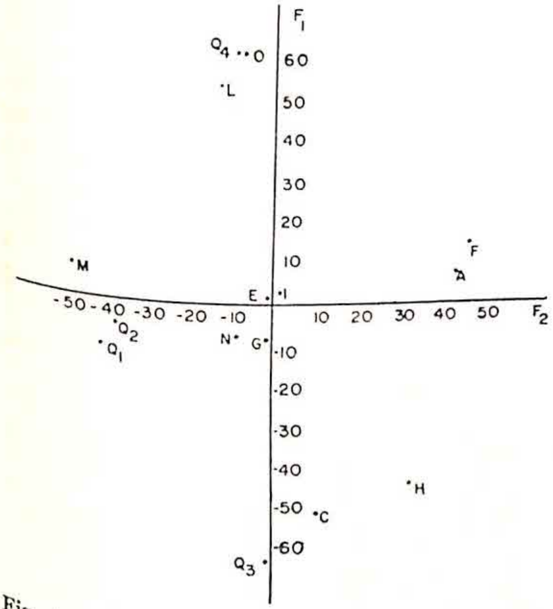


Fig. 1. Simple structure in the second-order realm: F_1 plotted against F_2.

The Second Order Factors

On the factor extraction from Table 2, carried out by the centroid method, Tucker's criterion (6) indicated four or five factors. Five were extracted, but subsequent rotation clearly collapsed one of them. One of the sources of confusion in second-order factor solutions, mentioned above, is that experimenters who rotate primaries for *oblique* simple structure nevertheless sometimes adopt *orthogonal*, unrotated, centroid interpretations for second-order factors. The well-known reasons for oblique simple structure in primaries (6, 30) actually apply just as forcibly to second-order factors. Simple structure may here be harder to determine without ambiguity, because of the fewer variables, but enough second-order studies now exist (12, 25, 30, 33) to show that simple structure commonly resides in the data and gives psychological meaning to factors thus uniquely determined. Accordingly, the second-order factors were rotated, and, after seven rotations, the structure shown in Table 3 was found. From the typical graph in Fig. 1 it will be evident that this is a very clear structure and practically unimprovable. (Possibly a 5° shift could still be made on each factor.) The number of variables in the hyperplane—

Table 3

Rotated Second-Order Factor Matrix

Primary factors	F_1	F_2	F_3	F_4
A	+.07	+.45	+.56	−.10
C	−.53	+.10	−.25	−.10
E	+.02	−.01	+.04	+.50
F	+.17	+.48	+.04	+.21
G	−.09	−.01	−.08	−.23
H	−.46	+.33	−.03	+.23
I	+.03	+.01	+.53	−.20
L	+.54	−.13	−.05	−.09
M	+.07	−.51	+.13	+.04
N	−.08	−.09	−.48	+.34
O	+.62	−.07	+.16	−.05
Q_1	−.12	−.42	+.05	+.55
Q_2	−.06	−.38	+.17	−.08
Q_3	−.66	−.01	−.29	−.02
Q_4	+.62	−.09	−.09	+.00

Note.—The unrotated matrix (the quartimaxed matrix, V_{oq}, from which the transformation matrix operates) and the transformation matrix by which this table was obtained have been deposited with the American Documentation Institute. Order Document No. 4969 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Table 4

Correlations of Second-Order Factors
(Simple structure after seven rotations)

	C^{-1}			
Factor	1	2	3	4
1	1.02			
2	.02	1.21		
3	.15	−.29	1.09	
4	.02	.43	−.13	1.14

Ednl. Psy. Research
TRAINING COLLEGE
Dated
No

slightly more than half—is about that found in most satisfactory first-order structures.

The intercorrelations of second-order factors $(D(\lambda'\lambda)^{-1}D)$, derived from the transformation (Lambda) matrix are shown in Table 4.

The chief indicators of the nature of Factors 1 through 4 in Table 4 will now be set out verbally, a contingent title being given to each factor.

Symbol	*Factor 1. Anxiety-vs.-Dynamic Integration*	*Loading*
$Q_3(-)$	Lack of will control[1]	.66
$O(+)$	Insecurity or free-floating anxiety	.62
$Q_4(+)$	Nervous tensions (somatic anxiety)	.62
$L(+)$	Paranoid trend	.54
$C(-)$	Lack of ego strength	.53

This factor has clearly brought together the "pathological" primaries. Before turning to interpretation one may point out that the evidence on second-order factors among behavior-rated personality primaries, which is published synchronously (12) with this article, shows a factor similar in all respects except that Factor I, called "Premsia" (12), is included there. Three theories compete for acceptance here: first, that it is a factor of general anxiety; second, a factor of general neuroticism; and third, that it is the different form of personality disorganization associated with psychoticism (20): see also Factor U.I. 25 in (7). The last has too many speculative elements and the first is somewhat weakened by the absence of the third main form of anxiety—desurgency ($F -$) or anxiety depression. On the other hand, it is striking that this set of five primary factors is identical with that found to correlate substantially with (and *only* with) the objective test factor (7) of Anxiety (indexed U.I. 24). (Both patterns also show some low H factor.) Nevertheless, considering the immediate nature of the above primary factors alone, one might incline to an interpretation of the second-order factor as Dynamic Integration-vs.-Neuroticism. There stands in the way of this interpretation Eysenck's interpretation of U.I. 23 as *the* neuroticism factor (20), an obstruction qualified somewhat by our reinterpretation of U.I. 23 as reduced neural reserves, rather than neurotic conflict per se. On the other hand, U.I. 24, with which the present second-order factor apparently identifies itself, is better described as anxiety than as conflict. Accordingly, since reliable description is more important than interpretation at this stage, it would seem best to compromise by naming the present second-order questionnaire factor *Dynamic Integration-vs.-Anxiety*.

Symbol	*Factor 2. Extraversion-vs.-Introversion*	*Loading*
$M(-)$	Practical concern or praxernia (12) (vs. Bohemianism or alaxia [12])	.51
$F(+)$	Surgency	.48
$A(+)$	Cyclothymia	.45
$Q_1(-)$	Conservatism	.42
$Q_2(-)$	Lack of self-sufficiency	.38
$H(+)$	Adventurous cyclothymia or parmia (12)	.33

Since the parallel study in rated behavior factors is only now appearing (12), it may be stated for the reader's information that a second-order factor (F_4) resembling this is found therein, combining high F and $M(-)$ with high D (not entering into questionnaire experiments). There, E factor, dominance, also enters in, and it is noteworthy that an alternative rotation considered here for Factor 2 on 4, reducing their correlation, would also bring it in here. A second confirmation of the pattern is found in that U.I. 32 among objective tests correlates significantly ($p = .01$) with Factors F, A, Q_2, H and E; though M and Q_1 are not as significant. These are thus, once more, quite independent indications of the same pattern.

This pattern has pulled together two primary factors which have been long recognized as cyclothyme-schizothyme in character, but it emphatically adds others which might go under the broader concept of extraversion—such as surgency, dominance, practical concern and lack of self-sufficiency. Along with conservatism these point to a person closely identified with the group life—"a good fellow"—as opposed to the independent recluse. The relatively low loadings of the more hereditarily determined H and A factors (14) suggest that Factor 2 is an environmentally determined pattern. There now seems ground for recognizing extraversion-introversion as something more than a correlation cluster or surface trait, and perhaps for specializing the term for this measurable second-order factor.

---

[1] The titles used are those of the 16 P.F. Handbook (15). To make easy, consistent reading, all loadings have been made positive, by reflection, in these factor lists. The reflection is indicated by the sign on the left and the change of title to the opposite pole, e.g., "Lack of. . . ."

Symbol	*Factor 3. Cyclothyme-vs.-Schizothyme Constitution*	*Loading*
$A(+)$	Cyclothymia	.56
$I(+)$	Emotional sensitivity or premsia (12)	.53
$N(-)$	Lack of sophistication	.48
$Q_3(-)$	Lack of will control	.29
$C(-)$	Lack of ego strength	.25

Again, a somewhat similar factor is found in the second-order among behavior-rated factors, in No. 5 (12), which links cyclothymia with low ego strength. A constitutional, body-build hypothesis was there suggested for experiment. Although only preliminary results are available it seems also to be the pattern (high *A*, *I*, low *N*) distinguishing manic-depressives from normals (15), and has fair resemblance to that distinguishing women from men (15) (but women have higher manic-depression and lower schizophrenia rates, as is well known). A constitutional body-build factor connected with cyclothymia is hypothesized.

Symbol	*Factor 4. Unbroken Success-vs.-Frustration*	*Loading*
$Q_1(+)$	Radicalism	.55
$E(+)$	Dominance	.50
$N(+)$	Sophistication	.34
$G(-)$	Low super ego strength	.23

Again it is noteworthy that a second-order factor (Factor 2) among observer-rated behavior factors (12) presents a pattern agreeing excellently with this in the variables common to the two studies (*E*, *N*, and *G*). The present study adds Q_1 and the objective behavior study adds high intelligence. Possible hypotheses are "easy personal success" occasioned by good intelligence or good educational opportunity, or "social status," which is known to be correlated with intelligence and dominance (2), and could well be correlated with sophistication and radicalism, in young men. Incidentally, the reader is reminded of the evidence that "radicalism," as a factor, is more than an accident of political belief, and constitutes a general temperamental source trait (19, 29).

Doubtless a generous native endowment in abilities plays a large causative role in this syndrome. One is reminded of a famous ambassador's comment on a British Secretary of State, known for his radical innovations, dominance of the world scene, sophistication, and debauchery (low *G!*): "his capacity is on so high a level that he has no need to give up his pleasures in the slightest degree for any cause" (17, p. 478). But in the life-record factor (12) the intelligence loading is scarcely high enough to make it the essential root, and the loadings are similar to those for social status. Consequently it is safer to consider this as the organization within personality brought about by unbroken success, from whatever cause, as opposed to frustration and intellectual bafflement. Among students it should correlate negatively with age.

Summary

1. A questionnaire—the 16 P.F.—broadly based on the personality sphere and over 2,000 items, had its factor structure redetermined on a sample of 408 young men and women, with particular attention to exact rotation to the best simple structure.

2. The correlations thus determined, i.e., by inversion of the reference vector matrix, among the 15 confirmed primaries (intelligence was omitted) factored to four second-order, simple-structure factors.

3. These four factors have resemblances, ranging from fair to excellent, to four second-order factors found from life-record (behavior rating) primaries, and two of them have closely similar profiles to the correlations found between questionnaire primaries and two objective personality test factors. The latter seems to confirm the surmise, made on general principles, that objective test factors would tend to be on a different level from rating and questionnaire factors, and, generally, to be second-order to them.

4. The factors have been contingently named *Dynamic Integration-vs.-Anxiety* (Objective test factor: U.I. 24); *Extraversion-vs.-Introversion,* (Objective test factor: U.I. 32); *Cyclothyme-vs.-Schizothyme Constitution;* and *Unbroken Success-vs.-Frustration.*

There is a significant positive correlation here of the second and fourth, as also was found in behavior rating second-order factors (12). Definite hypotheses are proposed for their further experimental investigation. As a part of such investigation it is desirable to score individuals on the 16 P.F. not only for primaries but also for these four second-order factors (from the above loadings) to see if higher correlations with a specific criterion are associated with second-order than with first-order factors. *Dynamic Integration* scores should be of particular interest to the clinician.

Received January 24, 1956.

References

1. Bargman, R. Significanzuntersuchungen der Einfachen Struktur in der Faktoren-analyse. *Mitteil. Math. Statist.*, Sonderdruk. Physica-Verlag. Wurzburg, 1954.
2. Cattell, R. B. The cultural functions of social stratification: 1. Regarding the genetic bases of society. *J. soc. Psychol.*, 1945, **21**, 3–23.
3. Cattell, R. B. *The description and measurement of personality*. Yonkers, N. Y.: World Book Co., 1946.
4. Cattell, R. B. Oblique, second-order and co-operative factors in personality analysis. *J. gen. Psychol.*, 1947, **36**, 3–22.
5. Cattell, R. B. The main personality factors in questionnaire, self estimate material. *J. soc. Psychol.*, 1950, **31**, 3–38.
6. Cattell, R. B. *Factor analysis*. New York: Harper, 1952.
7. Cattell, R. B. The principal replicated factors discovered in objective personality test. *J. abnorm. soc. Psychol.*, 1955, **50**, 291–314.
8. Cattell, R. B. Psychiatric screening of flying personnel. Personality structure in objective tests —A study of 1000 Air Force students in basic pilot training. *USAF Sch. Aviat. Med. Proj. Rep.*, 1955, Proj. No. 21–0202–007 (Rep. No. 9).
9. Cattell, R. B. Occupational profiles on the I.P.A.T. 16 Personality Factor Questionnaire. *Occup. Psychol.*, 1956, **30**, 1–10.
10. Cattell, R. B. Validation and intensification of the sixteen personality factor questionnaire. *J. clin. Psychol.*, in press.
11. Cattell, R. B. A shortened "Basic English" version (Form C) of the 16 P.F. Questionnaire. *J. soc. Psychol.*, in press.
12. Cattell, R. B. *Personality and motivation structure and measurement*. Yonkers, N. Y.: World Book Co., in press.
13. Cattell, R. B., Beloff, J., Flint, D., & Gruen, W. The Junior Personality Quiz. Champaign, Ill.: Institute for Personality and Ability Testing, 1602 Coronado Dr., 1954.
14. Cattell, R. B., Blewett, D. B., & Beloff, J. R. The inheritance of personality. A multiple variance analysis determination of approximate nature-nurture ratios for primary personality factors in Q-data. *Amer. J. hum. Genet.*, 1955, **7**, 122–146.
15. Cattell, R. B., Saunders, D. R., & Stice, G. F. The Sixteen Personality Factor Questionnaire. Champaign, Ill.: I.P.A.T., 1602 Coronado Dr., 1949.
16. Cattell, R. B., & Stice, G. F. Four formulae for selecting leaders on the basis of personality. *Hum. Relat.*, 1954, **7**, 493–507.
17. Churchill, W. S. *Marlborough, his life and times*. Vol. IV. London: Harrap, 1938.
18. Dexter, E. S. Personality traits related to conservatism and radicalism. *Charact. & Pers.*, 1939, **7**, 230–237.
19. Drevdahl, J. F. Factors of importance for creativity. *J. clin. Psychol.*, 1956, **12**, 21–26.
20. Eysenck, H. J. *The structure of human personality*. London: Methuen, 1953.
21. Fisher, R. A. *Social selection of human fertility*. New York: Oxford Univer. Press, 1932.
22. French, J. The description of personality measurements in terms of rotated factors. Princeton, N. J.: Educational Testing Service, 1953.
23. Guilford, J. P., & Guilford, Ruth B. Personality factors D, R, T and A. *J. abnorm. soc. Psychol.*, 1939, **34**, 1–12.
24. Guilford, J. P., & Zimmerman, W. S. The Guilford-Zimmerman Temperament Survey. Beverly Hills, Calif.: Sheridan Supply Co., 1949.
25. Hsü, E. H. On the correlation between a variable and its super-factor. *J. Psychol.*, 1946, **22**, 89–92.
26. Lentz, F. F. Generality and specificity of conservatism-radicalism. *J. educ. Psychol.*, 1938, **29**, 540–546.
27. Lovell, C. A study of the factor structure of thirteen personality variables. *Educ. psychol. Measmt*, 1945, **5**, 335–350.
28. Neuhaus, G. O., & Wrigley, C. The quartimax method: an analytical approach to orthogonal simple structure. *Brit. J. statist. Psychol.*, 1954, **7**, 81–91.
29. Suhr. V. W. The Cattell 16 P.F. as a prognosticator of accident susceptibility. *Proc. Iowa Acad. Sci.*, 1953, **60**, 553–561.
30. Thurstone, L. L. *Multiple-factor analysis*. Chicago: Univer. of Chicago Press, 1947.
31. Thurstone, L. L. Thurstone Temperament Schedule. Chicago: Science Research Associates, 1949, 1950.
32. Thurstone, L. L., & Thurstone, Thelma G. Factorial studies of intelligence. *Psychometr. Monogr.*, 1941, No. 2.
33. Vernon, P. E. *The structure of human abilities*. London: Methuen, 1950.
34. Wright, S. Some psychological and physiological correlates of certain academic underachievers. Unpublished doctor's dissertation, Univer. of Chicago Library, 1955.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Factors, Response Bias, and the MMPI

Eugene H. Barnes
VA Hospital, Downey, Illinois

Certain deductions concerning the influence of response set on the MMPI were tested in a recent article by Barnes (1). These deductions were derived from Berg's deviation hypothesis (2) and other pertinent research. This hypothesis asserts: "Deviant response patterns tend to be general; hence those deviant behavior patterns which are significant for abnormality and thus regarded as symptoms, are associated with other deviant response patterns which are in noncritical areas of behavior and which are not regarded as symptoms of personality aberration" (2, p. 62). Since the scoring system of the individual form of the MMPI provides that only those items answered in the infrequent direction are recorded, the MMPI provided a convenient source of deviant responses as defined by Berg. In his article Barnes (1) presented correlations of the MMPI scales with the number of items answered atypically true and with the number answered atypically false. He found the tendency to deviate by answering false to be more highly correlated with the scales of the neurotic triad (*Hs, D, Hy*) and the tendency to deviate by answering true to be more highly correlated with the scales of the psychotic triad (*Pa, Pt, Sc*). Further, the two tendencies were not significantly correlated with one another.

A factor analysis of the MMPI scales based on the scores of 110 neuropsychiatric patients was reported by Wheeler, Little, and Lehner (6) in which two major factors and one minor factor were identified. Factor I, the "psychotic" factor, had its maximal loading on *Pt* and *Sc*. This factor also had a substantial negative loading on *K*. Factor II, the "neurotic" factor, had its maximal loadings on the scales of the neurotic triad.

The correlations reported by Barnes (1) are sufficiently similar to the factor loadings on the two major factors reported by Wheeler, Little, and Lehner (6) to suggest that the sets may be empirical correlates of the rationally derived factors. Specifically, it is hypothesized that, if included in the correlation matrix, the vector representing number of true answers would be approximately collinear with Factor I, while the number of atypical false answers would be approximately collinear with Factor II.

Procedure

A direct test of this hypothesis does not appear possible since different samples are involved. An indirect method is to assume the existence of a test whose vector is collinear with the reference vector, calculate the correlations of each test in the matrix with the hypothetical test, and compare these calculated estimates with the obtained correlations of the test under investigation. In order to make these calculations it is necessary only to know the communalities of the tests in the original matrix, the cosine of the angle each makes with the reference vector, and the communality of the hypothetical test (5). Refined estimates of the communalities of the tests in the matrix are provided by the factor analysis. The cosine of the angle each test makes with the reference vector is simply its factor loading divided by the square root of the communality. The communality of the hypothetical test must be estimated. In this predicament it seems logical to use the same approximation recommended for such estimates in the original factor problem, i.e., the highest correlation obtained between the suspected collinear test and any of the other

Table 1

Correlations of Atypical Responses and Hypothetical Tests with Scales of MMPI

Scale	Atypically false	Hypothetical Test II	Atypically true	Hypothetical Test I
L	.37	—	−.24	—
F	.43	—	.66	.64
K	.06	—	−.69	−.67
Hs	.56	.63	.52	.50
D	.73	.55	.34	.46
Hy	.59	.72	−.09	—
Pd	.28	—	.68	.67
Mf	.21	—	.28	.36
Pa	.37	—	.52	.60
Pt	.31	—	.84	.89
Sc	.36	—	.90	.88
Ma	.21	—	.50	.59

Table 2

Communalities and Reliabilities of MMPI Scales

	Communalities	Reliabilities	
Scale	Wheeler, Little, & Lehner (6)	Rosen (4)	Holzberg & Alessi (3)
L	.234	.62	.85
F	.700	.81	.93
K	.643	.65	—
Hs	.873	.85	.67
D	.741	.80	.80
Hy	.771	.88	.87
Pd	.551	.88	.52
Mf	.338	.64	.76
Pa	.780	.75	.78
Pt	.944	.80	.72
Sc	.878	.83	.89
Ma	.494	.56	.58

tests in the matrix. In order to obtain estimated correlations using this approximation it is merely necessary to compute the scalar product of the two vectors, which turns out algebraically to be the square root of the estimated communality multiplied by the factor loading of the test with which its correlation is to be inferred. The communality of the test suspected to be collinear with Factor I (atypical true answers) is estimated at .90, its correlation with *Sc*. The communality of atypical false answers is estimated at .73, its correlation with *D*.

Results

The correlations expected for the hypothetical collinear tests as compared with the correlations obtained by Barnes (1) are presented in Table 1. It will be noted that estimates are not made for all tests on both factors. This is because factor loadings were not reported on these tests by Wheeler, Little, and Lehner (6). It is presumed that these omitted factor loadings were not considered significant. This would be consistent with the low predicted correlations.

It is obvious that the predicted correlations for a hypothetical test collinear with Factor I are in very close agreement with the experimental results of correlating the number of atypically true answers with the scales of the MMPI. The agreement of the predicted and obtained correlations for the number of atypically false answers is not so close, nor yet does it seem so divergent as to refute the hypothesis. Unfortunately there does not seem to be any significance test which can be applied to the kind of data presented. The meaning of the results cannot be evaluated, therefore, with the precision which is desirable.

Discussion

The correlations between the number of items atypically answered true and the scales of the MMPI are cogently similar to the correlations predicted for the hypothetical test, despite the lack of the significance test. There would appear to be little risk involved in accepting this measure as a "pure factor test" of the "psychotic factor." Although the predicted correlations for the number of atypical false answers do not have such a striking similarity, it is clear that, if included in the test matrix, this measure would have a heavy loading on the "neurotic factor." For this reason scores based on these response sets should, in themselves, be of considerable interest. It should be remarked, however, that a measure of the "psychotic factor" cannot be accepted as a measure of psychotic tendencies without further experimental evidence. A similar and more emphatic caution applies to the "neurotic factor," and the tendency atypically to answer false.

In addition to their interest as measures, it would seem that they might have another value. A comparison of the communalities reported by Wheeler, Little, and Lehner (6) with reliability coefficients reported by Rosen (4) and by Holzberg and Alessi (3) suggests that a degree of specificity remains in some of the scales. This comparison appears in Table 2. Since these two factors accounted for most of the communality of the scales of the MMPI and, since the two response sets appear to be relatively pure measures of the factors, it would appear that some adjustment of the scoring should be possible to reveal the remaining specificity. Such an adjustment should considerably increase the diagnostic precision of the MMPI. It is not clear to the writer exactly how this adjustment could be accomplished, but it seems inescapable that some allowance for response set will be necessary in the interpretation of the MMPI.

It is possible that the method of analysis used in this study may recommend itself to other investigators. The technique may be of value to the test constructor seeking to develop "factor pure" tests as a check of sorts on his success. Further, it might be used as a method of selecting measures to be included in a matrix for factor analysis. It would seem particularly appropriate if a previous analysis has given suggestive but somewhat equivocal results.

Summary

Two major factors, a "neurotic" factor and a "psychotic" factor, were identified by Wheeler, Little, and Lehner (6) in a factor analysis of the scales of the MMPI. Hypothetical tests collinear with these factors were postulated and the correlations of these hypothetical tests were computed for each of the scales in the correlation matrix for which factor loadings were reported. These correlations were compared with those obtained by correlating the number of atypical true answers to the MMPI with the scales of the MMPI and those obtained by correlating the number of atypical false answers with the scales of the MMPI. It was concluded that the number of atypical true answers is a "pure factor test" of the "psychotic factor" and that the number of atypical false answers has a heavy loading on the "neurotic factor." It was also concluded that some allowance should be made in the interpretation of the MMPI to allow for the influence of response set.

Received February 6, 1956.

References

1. Barnes, E. H. Response bias and the MMPI. *J. consult. Psychol.,* 1956, **20**, 371–374.
2. Berg, I. A. Response bias and personality: The deviation hypothesis. *J. Psychol.,* 1955, **40**, 61–72.
3. Holzberg, J. D., & Alessi, S. Reliability of the shortened MMPI. *J consult. Psychol.,* 1949, **13**, 288–292.
4. Rosen, A. Test-retest stability of MMPI scales for a psychiatric population. *J. consult. Psychol.,* 1953, **17**, 217–221.
5. Thurstone, L. L. *Multiple-factor analysis.* Chicago: Univer. of Chicago Press, 1947.
6. Wheeler, W. M., Little, K. B., & Lehner, G. F. J. The internal structure of the MMPI. *J. consult. Psychol.,* 1951, **15**, 134–141.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

The Personality of Judges and Their Agreement with Experts in Judging Clinical Case Histories

A. W. Bendig[1]
University of Pittsburgh

Large individual differences among judges in assessing clinical case histories suggest that variables such as intellectual ability, educational level, interests, and personality traits may influence their judgments. Such factors may be more important than the formal characteristics of the judgmental situation: the type of clinical material judged or the method of recording the judgments. The present study was designed as a preliminary investigation of the possible relation between certain objective personality traits and judgmental ability.

Undergraduate college *S*s ($N = 96$) ranked for global adjustment level 10 short clinical case histories with approximately half of the *S*s ranking a set of 10 cases of adult clinic patients and the other half ranking 10 case histories of disturbed children. The *S*s also completed the 15-scale Edwards Personal Preference Schedule (1). The rankings of each *S* were correlated with similar rankings of the same cases by 6 psychiatrists (adult cases) or 7 clinical child psychologists (child cases) and the *S*s dichotomized on the basis of the magnitude of their agreement with these experts. The *S*s were further divided into two replicated groups and the high and low agreement *S*s of Group I ($N = 56$) compared on each of the 15 scales of the PPS by two criterion (case type and agreement group) analysis-of-variance procedures. Two of the PPS scales, Achievement and Exhibition, significantly (.01 level or better) discriminated between the high and low agreement groups, while the interaction of case type and agreement group was significant (.05 level) with the Order scale. The high agreement group was higher on both the Achievement and Exhibition scales, while the high agreement *S*s ranking the adult cases and the low agreement *S*s ranking the child cases scored higher on the Order scale. Similar analyses of variance were computed on these three scales for the replication Group II ($N = 40$) and none of the resulting F ratios approached significance. Again the high agreement *S*s showed high mean scores on the Achievement and Exhibition scales, but the differences were not significant.

The obvious conclusion of this study is that there are no consistent relationships between the scores of naive *S*s on the scales of Edwards' PPS and their agreement with experts in judging clinical case histories. However, this does not exclude the possibility that other personality measures might show some correlation with judgmental ability.

Brief Report.
Received July 6, 1956.

Reference

1. Edwards, A. L. *Manual for the Edwards Personal Preference Schedule* New York: Psychological Corp., 1954.

---

[1] An extended report of this study may be obtained without charge from A. W. Bendig, Dept. of Psychology, University of Pittsburgh, Pittsburgh 13, Pa., or for a fee from the American Documentation Institute. Order Document No. 5039 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Comparison of Two Methods of Electroshock in Their Effect on Cognitive Functions[1]

Barbara S. Kendall, Warren B. Mills,
Washington University School of Medicine

and Thomas Thale
St. Louis University School of Medicine

Electroconvulsive therapy (ECT) commonly produces temporary confusion and disturbance of attention and of memory in the treated patient (6), or, as Stone (10) points out, a general impairment of cognitive functions. These unfortunate effects may accumulate during the treatment period, but tend to disappear within two or three weeks after the final shock. It has usually been assumed that ECT by the Reiter apparatus, because a lesser amount of current is necessary to produce a seizure, results in less disturbance to cognitive functions than does the conventional method (1, 2, 12, 13). Yet there have been no well-controlled studies dealing with measured differences in the effect on "test ability" of these two methods.

The aim of the present study is to test the following hypotheses:

1. For the same therapeutic result ECT by the Reiter method causes a lesser amount of impairment of cognitive functions, demonstrable at the end of the shock series, than does ECT by the conventional method.
2. Patients shocked by the Reiter method show less residual cognitive disturbance two weeks after the final convulsion than do patients treated by the conventional method.
3. Two weeks after the last convulsion, score gains in cognitive tasks over preshock performance are greater and losses are less for patients treated by the Reiter method than for those treated by the conventional method.

Method

Amount and type of treatment. Two groups of patients treated at the Malcolm A. Bliss Hospital are compared with reference to cognitive changes induced by ECT; one group, in which all were treated with the Reiter type 46 E machine (which delivers pulsating "unidirectional" current), will be referred to as the "Reiter group"; the "Conventional group" consists of those receiving treatment with the Lektra machine model 302 (which puts out a single burst of unmodulated 60-cycle sine wave alternating current). In order to insure uniformity, all patients were treated with a bitemporal placement of the electrodes, and no patient was presedated. This procedure was the usual one followed by this hospital in administering ECT. Treatments were scheduled three times a week until a full course had been run. Since it had been the experience in this institution that 20 convulsive reactions on the Reiter machine were usually required to induce approximately the same therapeutic result as that induced by 12 convulsions on the conventional type machine, each patient in the Reiter group received 20 and each one in the Conventional group 12 treatments. In a few instances, because of mechanical and other difficulties, single interruptions of the treatment rhythm resulting in intervals of four to seven days (instead of the expected two to three days) were unavoidable. These occurred only once for any one patient, involved four individuals in the Conventional and nine in the Reiter group, and tended to be longer in the latter group. Later it will be shown that these single interruptions did not alleviate the detrimental effect of ECT. One patient in the Reiter group received her sixteenth treatment on a Medcraft (conventional) machine.

---

[1] The authors acknowledge their indebtedness to Dr. Goldine C. Gleser for help in planning the statistical treatment of the data and to Dr. Laverne C. Johnson for reading the manuscript and for making valuable suggestions. This research was done at the Malcolm A. Bliss Psychopathic Hospital.

Sampling. The two groups of 20 patients each were matched for age, education, and intelligence. The mean age of the Reiter group was 30.40 years with a range of 21 to 52 years and an *SD* of 9.15; the mean age of the Conventional group was 34.35 years (range 20 to 59 years and *SD* of 9.59). For the Reiter group the mean number of years of education was 9.70 with a range of 6.50 to 15.00 and an *SD* of 2.51; for the Conventional group, it was 10.08 with a range of 4.00 to 16.50 and an *SD* of 2.89. The Wechsler-Bellevue Vocabulary was used to obtain a stable measure which would approximate original intellectual endowment. Form I was given before treatment and Form II two weeks after the close of treatment; whichever rating was higher was used for this score. The mean weighted score of the Reiter group was 10.70 points with a range of 7 to 16 and an *SD* of 2.41; that for the Conventional group was 10.75 points, with a range of 4 to 16 and an *SD* of 2.79.

There were 9 men and 11 women in the Reiter group and 8 men and 12 women in the Conventional group. Psychiatric diagnoses for the Reiter group were distributed as follows: schizophrenic reaction, 15 (of which 4 were of schizo-affective type); paranoid state, 2; involutional psychotic reaction, 2; and manic-depressive reaction, depressive type, 1. For the Conventional group they were: schizophrenic reaction, 16 (of which 2 were of schizo-affective type); paranoid state, 1; involutional psychotic reaction, 1; manic-depressive reaction, depressive type, 1; and depressive reaction, 1. None of the patients had received any type of convulsive treatment within the previous year; no case with known brain damage was included.

It was not possible to match the two groups for severity of illness, and selection from this point of view was necessarily at random. However, certain rough checks may be applied. Duration of illness is not significantly different in the two groups: 9 patients in the Reiter and 8 in the Conventional group had been ill for one year or less; comparable figures when the criterion is two years are 12 and 10. An index of Wechsler-Bellevue scatter which Pascal and Zeaman (8) have shown to differentiate between degrees of severity of psychiatric illness yields means of 9.35 for the Reiter and 8.06 for the Conventional group; the difference is not significant (*t* equals 1.01).

Psychological tests. In order to measure the amount of postshock confusion two short test batteries were set up, to be administered on the day of the first treatment, on the day when half the course of electroshock was completed (in the case of the Reiter group on the tenth treatment day and for the Conventional group on the sixth day), and, finally, on the last day of treatment.

The first battery, consisting of short tests of Picture Absurdities (7), Substitution (7), Cancellation (Whipple), and Cube Imitation (Knox), was invariably given 30 minutes after the convulsion, and will be referred to as the Morning Battery. The second, used (with few exceptions) three to three and a half hours after the convulsion, will be called the Noon Battery; this included naming the months in reverse, subtracting serial 7's (9), and serial 3's, naming colors (9), duplicating from memory patterns of colored blocks (9), and identifying from a pack of 15 cards 5 which carried words just previously used as stimulus words in an opposites test (5). Two patients in the Conventional group, one an aggressive chronic paranoid and the other a catatonic in her first episode, refused to cooperate in the Noon Battery on the first shock day, but gave adequate cooperation three hours after the second treatment. Where possible, alternate forms of the tests of these batteries were arranged for use at the second and third sessions. No change seemed practicable in Months Reversed, Picture Absurdities, Cube Imitation, Color Naming, or the Ruesch Blocks; but slight changes were made by shifting the key in Substitution, crossing out a different letter in Cancellation, and beginning with a different number in Serial 7's and Serial 3's; for the Recognition test the subject was asked to find opposites for a new set of words. Thus, although transfer of training could not be eliminated, the influence of previous practice was minimized for half of the tasks without introducing, except in the case of Recognition, a real uncertainty about equivalence of difficulty.

In order to have a measure of the rate of recovery from the temporary intellectual inefficiency which is the immediate cumulative effect of ECT, the Wechsler Memory Scale Form I was administered to all patients immediately after the Noon Battery on the final shock day, and this was followed approximately two weeks later with Form II of the same scale. Longer-term losses and gains in intellectual efficiency which resulted from the treatment were evaluated by means of the Wechsler-Bellevue Scale. When possible, Form I was given before the first shock, and Form II 13 to 15 days after the last shock.

Results

Measures of Postshock Confusion

In the analysis of results *T* scores were computed, according to the method of McCall, for each subtest of the Morning and Noon Batteries, and these were based (with one exception) on all three performances of both groups in the same or equivalent tasks. The Recognition test was given after the first and final shocks only, and, consequently, 80 scores, instead of 120, were used in the calculations. It would not have been useful to score the subtests of the battery by already established norms (where these existed) because the scores earned after ECT tended to accumulate at the lower end of each distribution to an extent which precluded effective differentiation. It was also undesirable to normalize for each administration separately because the scores would have failed to show the extent of the changes which occurred in the course of treatment. The method chosen, i.e., the normalizing of scores of each subtest for all three administrations, does not represent an attempt to establish standards of performance, but is rather a means of providing a basis for comparisons between Reiter and Conventional groups and among the three administrations of the subtests.

In Table 1 mean scores for each group in each subtest administered on the first day of treatment are compared with mean scores for those given on the last day of treatment, and a cross comparison is made of the differences between gains and losses shown by both groups. Means were computed for midcourse measures, but, since they add little to the analysis, are not included in the table.

Changes induced by ECT. Test-retest changes are considered separately for each experimental group. Decrements significant at or beyond the 2% level of confidence occur for both groups in Substitution and Serial 7's, and at or beyond the 5% level in Cancellation. For one or the other group losses significant at or beyond the 1% level are found in Picture Absurdities, Serial 3's, and Recognition, and at the 5% level in Cube Imitation and Months Reversed. Thus 8 of the 10 subtests are shown to be more or less sensitive to the cognitive changes induced by ECT. In the case of Color Naming, however, the effect of previous practice appears to cancel out the damaging effect of ECT; and Block Design actually shows a gain, though for the Reiter only, which is significant at the 5% level.

Test of Hypothesis I. The difference in gain and loss from first to final shock is expressed as the average change for Reiter minus the average change for Conventional. A comparison of the two groups with regard to these score decrements or increments yields only one significant difference, that for Recognition. A possible reason why this test of memory and alertness made the differentiation where the other tests of postshock confusion failed to do so is that transfer of training is more successfully minimized for this task than for any of the others, partly because a new set of stimulus words is used in the retest, and partly because the midcourse retest is omitted. It may be that for the other subtests the effect of previous practice offset in some degree the effect of confusion, but in a test of new learning the Reiter group was better able to demonstrate superiority.

Memory Scale

Changes in first two weeks of recovery period. In contrast to the experience with tests used to measure changes taking place during the course of treatment, practice effects in the case of the Wechsler Memory Scale, which is used as a measure of recovery, tend to inflate the retest scores and make the differences artificially greater. In Table 2 the mean Memory Quotients on the last day of ECT are compared with results of a second administration two weeks later. For both groups the test-retest gains are significant beyond the 1% level of confidence.

Further test of Hypothesis I. While the majority of the tests of postshock confusion

Table 1

Measures of Postshock Confusion: Means, Sigmas, and Measures of Difference in Reiter (R) and Conventional (C) Groups (*T* Scores)

Test	First shock (1)		Final shock (3)		Diff. 1−3	*t*	D_R−D_C	*t*
	Mean	Sigma	Mean	Sigma				
Morning battery								
Picture Abs.								
R	52.60	6.04	46.20	9.35	6.40	2.86***	2.10	.63
C	52.10	7.30	47.80	9.28	4.30	1.76		
Substitution								
R	54.60	11.04	45.85	9.52	8.55	3.88***	1.25	.35
C	53.90	9.43	46.40	8.27	7.50	2.72**		
Cancellation								
R	52.30	9.23	47.25	9.71	5.05	2.12*	−.40	.13
C	53.05	9.34	47.60	9.91	5.45	2.75**		
Cube Imitation								
R	54.40	8.61	50.05	8.86	4.35	2.21*	.85	.23
C	49.40	10.45	45.90	9.33	3.50	1.13		
Noon battery								
Months Rev.								
R	53.75	10.79	48.15	11.50	5.60	2.32*	2.60	.92
C	50.95	9.28	47.95	8.64	3.00	2.02		
Serial 7's								
R	54.95	10.78	48.30	9.18	6.65	4.21***	1.65	.84
C	52.00	10.27	47.00	7.93	5.00	4.23***		
Serial 3's								
R	50.60	12.69	49.10	8.77	1.50	.59	−2.20	.79
C	51.35	9.63	47.65	8.56	3.70	3.29***		
Color Naming								
R	51.05	13.00	48.63	10.62	2.42	1.08	−.58	.17
C	51.15	10.25	48.15	8.26	3.00	1.18		
Blocks								
R	46.55	10.50	52.90	9.85	−6.35	2.42*	−3.75	1.15
C	47.55	9.43	50.15	3.44	−2.60	1.35		
Recognition								
R	51.00	6.98	53.60	9.26	−2.60	1.01	−9.80	2.72***
C	50.70	9.37	43.50	7.10	7.20	2.86***		

* Significant at the 5% level of confidence.
** Significant at the 2% level of confidence.
*** Significant at the 1% level of confidence.
Note.—Means are rounded. In all cases higher scores indicate more favorable performance.

fail to bring out clear-cut differences between the effects of the two methods of administering ECT, results for the Wechsler Scale on the final shock day give striking evidence that the Reiter group shows less cognitive disturbance. Significance beyond the 1% level of confidence is indicated.

A comparison of subtest scores reveals that the Reiter group surpasses the Conventional in several measures of new learning, i.e., in Logical Memory, Memory for Designs, and the "hard" items of Associate Learning; differences of 2.73, 1.90, and 1.05 are significant (in the first instance at the 1% level, and for the last two at the 5% level of confidence); *t* is, respectively, 2.90, 2.08, and 2.10. No other difference in subtest scores between the two groups (except in Orientation) approaches statistical significance.

Because the two groups are matched for intelligence, age, and education it is assumed that differences found at the first administration of the Memory Scale reflect differences in the amount of cognitive impairment induced by the ECT series. An attempt to match the groups for the memory factor by

Table 2

Measures of Recovery from Postshock Damage: Means and Sigmas in Reiter (R) and Conventional (C) Groups

	Pretreatment (I)		Final shock (I)		After 2 wks. (II)	
Test	Mean	Sigma	Mean	Sigma	Mean	Sigma
Wechsler MQ						
R			76.95	11.36	90.60	20.28
C			66.55	8.86	81.90	16.95
Wechsler-Bellevue IQ						
R(*N* 20)	98.85	12.68			95.05	13.26
C(*N* 16)	103.31	12.27			99.50	10.16
C(*N* 17)	103.47	11.92				
C(*N* 19)					96.42	13.30

	Differences and their significance							
	I−II	*t*	RI−CI	*t*	RII−CII	*t*	$D_R - D_C$	*t*
Wechsler MQ								
R	−13.65	3.47**	10.40	3.15**	8.70	1.44	−1.70	.32
C	−15.35	4.39**						
Wechsler-Bellevue IQ								
R	3.80	1.89	−4.62	1.11	−1.37	.31	−0.01	.01
C	3.81	2.19*						

* Significant at the 5% level of confidence.
** Significant at the 1% level of confidence.

Note.—The Wechsler-Bellevue data for the Conventional group are incomplete; one patient was mute and two had been given only four subtests prior to shock; a fourth refused the posttest Performance Scale. Differences are calculated for the varying numbers of patients who gave full records in the different situations.

using the scale as a pretest or as a posttest would have been illusory. The groups comprised a variety of psychiatric conditions, and the schizophrenic patients were in different stages of the illness from acute to chronic. Any pretest, other than vocabulary, would have fallen far short of representing inherent intellectual endowment. The remission of the psychosis brought about by ECT should be maximal immediately after the conclusion of the series, and its effect more nearly equal for all patients than at any other time, but this effect was essentially nullified for cognitive tests by the intellectual impairment which all the experimental patients suffered. Two weeks later, when the retest was made, some patients had begun to lose the benefits accruing from the treatment while others maintained their gains. In either situation the establishment of approximate equivalence for the two groups in a sensitive measure was not possible. Matching for the more stable attributes of vocabulary, age, and education, though not ideal, was the most that could be achieved.

Test of Hypothesis II. Two weeks after the last treatment the difference between the means of the two groups in the Memory Scale, though not greatly diminished, is no longer statistically significant, and the difference in the extent of test-retest changes, i.e., in mean gains or losses, is negligible. A wider spread of scores, shown in what is almost a doubling of the standard deviations, reflects a variability in immediate outcome. Stone (11) has found the Wechsler Memory Scale to be "remarkably sensitive" to the temporary intellectual impairment associated with ECT. It seems also to be sensitive to changes which, in those cases which derive no lasting benefit, are associated with a reactivation of the psychosis for which a temporary remission has been effected by ECT. Ten patients of the Reiter and 12 of the Conventional group had improved sufficiently to be slated for an early discharge. Among the remaining patients, three were signed out against advice, a few were discharged later, but others required long-term care. The mean gains for the "good" outcome group—20.00

—and for the "poor" outcome group—7.78—differ by 12.22 points, and this difference is significant at the 2% level.

Effect of interrupted treatment. Since the amount of gain shown in the final administration of the Memory Scale might be thought of as a rough measure of the extent of impairment of brain function which resulted from the full course of ECT—impairment which was actually reversed in the two-week period following treatment—an attempt is made to use this measure to determine whether the patients whose courses of treatment were interrupted had suffered less temporary detrimental effect as a consequence. Eleven patients in the Reiter group who had uninterrupted treatment were compared with 9 in the same group for whom 4-, 5-, and 7-day intervals occurred once for each case. The mean gain for the former group was 5.45 and for the latter 23.66; the difference is significant at the 2% level ($t = 2.67$), but is not in the anticipated direction. This difference is not explained by differences in intelligence level, or by a significant inferiority in the "continuous treatment" group immediately after the course of ECT. Although it appears that improvement in the first two weeks following shock was somewhat related to intelligence (the product-moment correlation between gains and Vocabulary score being .40, significant at the 1% level of confidence), differences in Vocabulary and in the preshock Wechsler-Bellevue IQ between patients of the Reiter group having interrupted treatment and those having continuous treatment (.64 and 8.66, respectively) are not significant ($t = .54$ and 1.49). For the memory quotient earned on the final shock day the mean of the "interrupted treatment". group surpassed that of the "continuous treatment" group by only 3.12; the difference is not significant ($t = .55$). We are at a loss to explain why a greater amount of reversible cognitive disturbance should occur to those patients whose treatment, because of one unduly long rest interval, was spread over a longer period of time. There is no evidence that these brief interruptions operated to alleviate the disrupting effect of treatment.

Wechsler-Bellevue Scale

Changes pre- to postshock. Also in Table 2 are shown mean IQs and their standard deviations for two administrations of the Wechsler-Bellevue Scale; Form I was given before the start of ECT and Form II a fortnight after its completion. Though in individual cases the changes may be marked (the range being 30 IQ points lost to 9 IQ points gained), for each group as a whole a slight decrement is found on retest, which is significant, at the 5% level, in the case of the Conventional group and approaches significance for the Reiter group. The experience of Gerboth (3) and of Hayes and Schneider (4) suggests that with normal subjects a negligible gain results when Form II follows Form I. Here the expectation would be that remitting psychosis, present in more than half the cases, would bring about an increase in scores, but, presumably because after two weeks there is still residual impairment of cognitive functions, this effect is not found. Confirming evidence that intellectual loss might still be present on retest is brought out in one qualitative feature, a difficulty with word finding, which occurred in a few instances. In responding to Picture Completion two weeks after the end of the shock series, four patients, though they indicated by pointing that they understood what was missing in item 3, were unable to recall the word *hinge*, and five others recalled it only after a marked hesitation.

Test of Hypothesis III. The difference between the experimental groups in average changes (gains or losses) in Wechsler-Bellevue IQ from first to final administration is negligible. Therefore the hypothesis that pre- to postshock average gains would be greater, and average losses less, for the Reiter than for the Conventional group is not sustained.

Summary

In order to compare the detrimental effect on cognitive functions of two different methods of administering ECT, two groups of 20 patients each, matched for age, education, and intelligence, were subjected to ECT, the first group to 20 convulsions by the Reiter method, and the other to 12 convulsions by the Conventional method. Various psychological measures were used before shock, during the course of shock, and two weeks postshock.

At the close of the treatment period less cognitive disturbance, as measured by certain tests involving new learning, had occurred in the Reiter group, but differences were negligible for other tasks. In the two-week postshock period gains in the Wechsler Memory Scale were approximately equal for the two groups. Comparison of pre- to postshock average changes in the Wechsler-Bellevue Scale gave no advantage to either group.

Received January 24, 1956.

References

1. Friedman, E. Unidirectional electro-stimulated convulsive therapy. *J. nerv. ment. Dis.*, 1949, **109**, 540–549.
2. Friedman, E., & Wilcox, P. H. Electrostimulated convulsive dazes in intact humans by means of unidirectional currents. *J. nerv. ment. Dis.*, 1942, **90**, 56–63.
3. Gerboth, Renate. A study of two forms of the Wechsler-Bellevue Scale. *J. consult. Psychol.*, 1950, **14**, 365–370.
4. Hayes, W., & Schneider, B. A test-retest evaluation of the Wechsler Forms I and II with mental defectives. *J. clin. Psychol.*, 1951, **7**, 140–143.
5. Jellinek, E. M., & Bolles, Marjorie. The effect of bromides on certain psychological test performances. *J. Psychol.*, 1942, **14**, 277–293.
6. Kalinowsky, L. B., & Hoch, P. H. *Shock treatments, psychosurgery, and other somatic procedures in psychiatry.* (2nd Ed.) New York: Grune & Stratton, 1952.
7. *Kuhlmann-Anderson Intelligence Tests.* (5th Ed.) Minneapolis: Educational Test Bureau, 1942.
8. Pascal, G. R., & Zeaman, Jean B. A note on the validity of Wechsler-Bellevue scatter. *Amer. J. Psychiat.*, 1949, **105**, 840–842.
9. Ruesch, J. Intellectual impairment in head injuries. *Amer. J. Psychiat.*, 1943, **100**, 480–496.
10. Stone, C. P. Losses and gains in cognitive functions as related to electro-convulsive shocks. *J. abnorm. soc. Psychol.*, 1947, **42**, 206–214.
11. Stone, C. P., Girdner, J., & Albrecht, Ruth. An alternate form of the Wechsler Memory Scale. *J. Psychol.*, 1946, **22**, 199–206.
12. Summerskill, J., Seeman, W., & Meals, D. W. An evaluation of postelectroshock confusion with the Reiter apparatus. *Amer. J. Psychiat.*, 1952, **108**, 835–838.
13. Wilcox, Katherine W. Confusion and therapy in electroconvulsive treatment. *Conf. Neurol.*, 1954, **14**, 318–326.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Group Psychotherapy with Chronic Psychotic Soiling Patients[1]

John E. Tucker
VA Hospital, Albany, N. Y.

This study was designed to test the effectiveness of group psychotherapy under controlled experimental conditions with an objective criterion. Twenty male psychotic *S*s without detectable organic bases for their soiling habits were selected from a ward of 38 soilers. A soiler was defined as a psychotic patient with urinary or fecal incontinence.

The frequency of soiling for the 20 *S*s was measured in a 15-day pre-experimental period using standard tally sheets kept by trained ward nurses. The 20 *S*s were given a brief test of mental alertness which gave a quantitative score. The *S*s' ages were taken from medical records. The *S*s were paired by trial and error into 2 groups of 10 each so that the means and *SD*s of their ages, soiling frequencies, and mental-alertness test scores were not statistically different between the 2 groups (critical ratios for M_d: .590, .614, .870). The *S*s were diagnosed chronic psychotics. One group was randomly designated the control group and did not have group therapy. The other group was the experimental group and was given 45 hours of group therapy.

Therapy was oriented toward increasing verbal communication. The *S*s were regressed and disorganized. Transference was fostered by feeding the *S*s sweets. No overt changes were noted in *S*s' psychotic behavior but they showed friendliness toward the therapist and did not resist group meetings. The *S*s achieved some psychological group structure as shown by consistent choice of seats, the formation of submissive-dominant relationships, and some role-taking.

During the 9-week experimental period the 20 *S*s stayed on their usual ward and did not have any change in their living routine except for group therapy for the 10 experimental *S*s. The identities of the 10 controls were concealed from ward personnel. The tally of soiling was kept over the 9-week period for both groups. At the end of the period the pre-experimental soiling data were compared with postexperimental and mean gain differences were analyzed. Controls decreased 1.1 in their daily soiling average. Experimentals decreased 6.7 in theirs. The 5.6 difference favoring the experimental group mean gain was statistically significant, p less than .05.

The selection and matching procedures embodied in the design encourage the hypothesis that the greater reduction in soiling behavior in the experimental *S*s resulted from group therapy. Possibly, group therapy was a sufficiently potent technique to modify early learned behavior patterns such as toilet habits. Use of group therapy with other behavior pattern disturbances found in psychiatric hospitals such as assaultiveness or elopement suggests itself.

Brief Report.
Received July 24, 1956.

---

[1] An extended report of this study may be obtained without charge from John E. Tucker, VA Hospital, Albany, New York, or for a fee from the American Documentation Institute. Order Document No. 5040 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

The Short-Term Effects of Chlorpromazine on Psychiatric Patients

William D. Winter and Wilbur K. Frederickson
Hospital for Mental Diseases, Milwaukee, Wisconsin [1]

The increasing use of chlorpromazine and other tranquilizing drugs in the treatment of mental disorders poses practical and theoretical problems for the clinical psychologist working in this area. Experience with these drugs has demonstrated their efficacy in making disturbed patients more manageable and reducing the need for restraints, sedative packs, and seclusion (5). Used in conjunction with psychotherapy, the drugs have replaced to a large extent the more drastic "shock" therapies as the treatment of choice. The exact physiological and psychological action produced by these drugs is still largely unknown. We see clinically a reduction of anxiety, agitation, and emotional lability in the patient which makes him more amenable to psychotherapy. It is still undetermined whether these reactions indicate that basic personality changes are being effected in an individual or whether the drugs only enable the patient to work out his problems more effectively by lessening the confusion and intense emotionality set off by unbearable conflicts.

One purpose of the present study will be a preliminary investigation into the nature of personality changes produced under chlorpromazine, as measured by ward behavior and psychological tests. Our first hypothesis is that within a week's time no significant personality trait changes, as measured by the test used in this study (MMPI), would occur as a direct result of chlorpromazine treatment. We, however, predict an improvement in overt ward behavior, related to a decrease in subjective anxiety.

Another general problem investigated, a practical one for psychologists, concerns the validity of administering tests involving visual-motor coordination to patients undergoing chlorpromazine treatment. In this hospital most patients are routinely put on some form of drug therapy a few days after their admission. Physical symptoms due to the physiological action of the drugs occur shortly after the patients begin this therapy. These symptoms usually include blurred vision, difficulty in coordination, and dizziness. A test involving visual-motor coordination, such as the Bender-Gestalt test, can be expected to be particularly vulnerable to misinterpretation if the side reactions to the drugs significantly affect test performance. Our testing experience with patients undergoing drug therapy led us to predict that tests involving a visual-motor task would be contaminated by the physical side effects of the drug.

Methodology

The subjects used in this study were 59 female patients newly admitted to the Milwaukee County Hospital for Mental Diseases. These patients represented a variety of diagnostic categories, but all those with known or suspected organic impairment or those in need of emergency care were excluded from the sample. The age range of the subjects was from 18 to 45 years, with a median age of 33.5. The median educational level was 11.5 years of schooling, and the median mental age was 15.5, as estimated from the Shipley-Hartford scale.

---

[1] The authors wish to express their appreciation of the assistance given them in carrying out this study by Christine Ivanov, M.D., Donald Schaefer, M.D., Robert Lichtenstein, M.D., and the nursing staff of this hospital.

The test battery included the Bender-Gestalt test, the group form of the MMPI, and the Shipley-Hartford scale. The Gardner Behavior Chart (8) was used to estimate adjustment to ward routine. A check list showing the deviation of the patient's drawings from the stimulus figures was used to score the Bender-Gestalt test. This check list was developed by the authors at the suggestion of Max Hutt. This list incorporates many of the items in Pascal and Suttell's scoring system (6), and also includes deviations such as constriction and overexpansiveness, which are found in Hutt's criteria (4). Unlike Pascal's system, differential weighting of specific deviations was not attempted. Scoring reliability for this check list was found to be .91. The scoring of the MMPI followed the practice of adding the designated percentage of *K* to the necessary diagnostic scales (3).

As patients were admitted to the hospital, they were randomly assigned to two groups. Both groups received placebos, which were identical to the chlorpromazine tablets in size and color, for three days. At the end of this time, each patient was interviewed and tested with the Bender-Gestalt test and the MMPI, and a nurse who was familiar with the patient filled out the Gardner Rating Chart on her. On the fourth day, the patients in the Drug group began receiving 50 mg. tablets of chlorpromazine (thorazine) 4 times a day, and the Placebo group continued getting placebos. Seven days later, both groups were retested on the Bender and MMPI and another Gardner Rating Chart was filled out. The Shipley-Hartford scale was also administered at this time. At no time did the physicians, nurses, or psychologist doing the testing know which of the two coded treatment routines involved the use of chlorpromazine. The Bender records were also coded so that the scorers would not know whose test was being scored or whether the protocol was from the first or second test administration.

The two groups were not significantly different on the variables of age, education, intelligence, number of admissions to the hospital, or diagnostic category.

Table 1

Test-Retest Reliability Coefficients

Measure	*N*	*r*
Ward Ratings	51	.86
Bender-Gestalt	44	.80
MMPI *T* Scores		
L	31	.78
F	31	.76
K	31	.65
Hs	31	.76
D	31	.79
Hy	31	.66
Pd	31	.71
Mf	31	.63
Pa	31	.75
Pt	31	.86
Sc	31	.77
Ma	31	.65

Results

In Table 1 are listed the test-retest reliability coefficients for each of our measures with the time interval of one week between tests. These correlations are based on the scores of the drug and placebo groups combined, and are similar to the average reliabilities found in previous studies (1, 3, 7). The median reliability for the placebo group alone was .75; for the drug group, .745.

The changes occurring in the test scores over a one-week period are given in Table 2

Table 2

Significance of Changes Occurring in a One-Week Period for Entire Sample

Measure	*N*	Mean score on first testing	Mean change	*p*
Ward Ratings	51	36.8	+3.1	<.01
Bender-Gestalt	44	27.5	−2.6	<.01
MMPI *T* Scores				
L	31	57.2	+1.7	<.20
F	31	60.1	−1.9	<.20
K	31	59.4	+2.5	<.10
Hs	31	58.6	−1.2	<.40
D	31	68.2	−3.2	<.20
Hy	31	65.2	−2.8	<.20
Pd	31	72.0	−4.7	<.02
Mf	31	54.7	−1.5	<.30
Pa	31	59.0	−1.9	<.30
Pt	31	63.0	−3.9	<.02
Sc	31	65.2	−4.1	<.02
Ma	31	55.3	−2.7	<.20

for the entire sample. Change was computed by subtracting the first test score from the second test score of each subject, and the significance of this difference was determined by the use of two-tailed t tests for correlated data. As can be seen from the table, our subjects as a group showed significant improvement in ward behavior and Bender-Gestalt scores, and a significant decrease on the Psychopathic Deviate, Psychasthenia, and Schizophrenia scales of the MMPI. They also showed a tendency toward a higher K score on the MMPI at the time of their second testing.

Inspection of the data revealed that the two groups differed significantly on several MMPI scales on the original testing, therefore, the drug and placebo patients were matched as closely as possible on the basis of their initial scores on each measure, in order to determine which group changed more in one week's time. Comparison of the mean changes of the two groups by use of the t test for uncorrelated means revealed only three variables in which the differences approached significance at between the .10 and .05 level of confidence. The patients receiving drugs improved more than those taking placebos on the Ward Ratings, and the placebo group improved more on the Bender. The drug group showed significantly higher scores on the K scale of the MMPI. The number of subjects compared on Ward Ratings was 42, on the Bender-Gestalt 32, and on the K scale 22.

Discussion

Analysis of the data indicates that in a period of one week the patients as a group showed small but statistically significant changes in Ward Behavior ratings, Bender-Gestalt scores, and on several MMPI scales. This is congruent with our own clinical experience with acutely disturbed patients, and with the result of other studies (7); the severity of their symptoms often decreases markedly after only a few days of hospitalization. The problem of the depth and permanence of these changes cannot be completely solved. It should be pointed out that our data are limited and we cannot precisely evaluate the effects of practice, the lack of independence of the MMPI scales, the short time interval, and the use of only a moderate drug dosage.

When we compare the mean changes of the drug group with those of the placebo group, we do not find any outstanding differences. There is a trend, however, for the drug group to show better ward adjustment, poorer coordination on the Bender-Gestalt, and a higher K score. The first two findings are in the predicted direction. The use of chlorpromazine reduces subjective anxiety and excessive emotional lability, which makes the patient easier to get along with and more approachable. At the same time, particularly in the early phases of drug therapy, the patient suffers from many side reactions such as dizziness, blurred eyesight, poor coordination, and physical unrest, all of which could easily affect performance on the Bender-Gestalt test.

Apparently none of the usual diagnostic scales on the MMPI were differentially affected by the drug. The test-retest reliability figures for the two groups on the MMPI are similar to those cited in the MMPI manual and in other studies (3, 7). This suggests that the use of chlorpromazine therapy with the patients resulted mainly in superficial behavioral improvement which may have involved a suppression of symptoms rather than deep personality changes. There could, of course, be an increase in ego strength which may lead to eventual major changes in personality structure, but the data shed no light on the question. It is interesting that other studies in this area have shown little or no change in psychological tests after drug therapy despite such overt changes in behavior as to markedly lessen the need for sedative packs, electric shock, seclusion, etc. (2, 5).

A comment should be made concerning the need for using control groups when doing research in this area. As mentioned above, a certain percentage of acutely ill patients, even when not given any special kind of treatment except for routine hospitalization, tend to get better. When control groups cannot be used, it is easy to attribute this improvement to whatever therapy is being administered. In one case a physician hesitated about putting a patient on the project because she was quite disturbed and he wanted her to receive electric shock. She was nevertheless placed on the project where she made spectacular improvement in her ward behavior. The physician was reassured that the patient was at least getting some drug therapy, but a later check of the records showed that the woman was in the placebo group. Thus the physician was falsely attributing her spontaneous (i.e., of unknown cause) recovery to the action of the drug.

In interviews with patients we were made aware of the effect of group suggestion. Many patients entered the hospital after reading optimistic newspaper reports of the "wonder drugs," and every patient was exposed to the enthusiasm of the medical staff and to the comments of other patients. Sometimes a patient in the placebo group would report not only a lessening of tension due to the supposed drug she was receiving, but would also dwell on the physical discomfort the medication caused.

Improvement in both groups varied considerably from individual to individual. For example, the change in ward behavior rating on the placebo group ranged from − 12 to − 15, and for the drug group from − 4 to − 14.

The drug, then, can be said to cause some apparent changes for the better in a seven-day period, but it is not possible to say how central these changes are or how deep they go. Individual differences are great, and the therapeutic effects may sometimes be exaggerated in uncontrolled research by neglecting the factors of spontaneous improvement, suggestion, and biased observation.

Summary

A group of acutely disturbed female patients was tested before and after a week of chlorpromazine therapy, and the changes in test scores were compared with those of a control group which received only placebos. The combined groups showed small but significant changes in Ward Ratings, Bender-Gestalt test scores, and on the *Pt*, *Pd*, *Sc*, and *K* scales of the MMPI. When patients were matched on their original score for each measure, the data showed a trend of the drug group to improve more in Ward Ratings and increase more on *K*, but to do less well on the Bender-Gestalt. Our conclusions are that, in general, acutely disturbed patients tend to improve after a few days of hospitalization and that chlorpromazine speeds behavioral adjustment to ward routine, but that it causes some disruption of coordination, especially during the so-called "turmoil phase" at the beginning of treatment. Evaluation of psychological tests involving visual-motor coordination, such as the Bender-Gestalt, should, therefore, take into account the effects of concurrent drug therapy.

Received April 9, 1956.

References

1. Addington, M. A note on the Pascal and Suttell scoring system of the Bender-Gestalt test. *J. clin. Psychol.*, 1952, **8**, 312–313.
2. Cowden, R. C., Zax, M., & Sproles, J. A. Group psychotherapy in conjunction with a physical treatment. *J. clin. Psychol.*, 1956, **12**, 53–56.
3. Hathaway, S., & McKinley, J. C. *Minnesota Multiphasic Personality Inventory: Manual.* New York: Psychological Corp., 1951.
4. Hutt, M. L. Revised Bender Visual-Motor Gestalt Test. In A. Weider (Ed.), *Contributions toward medical psychology*. New York: Ronald, 1953. Pp. 660–688.
5. Kasak, M., Schaefer, D., & Winter, W. D. The use of chlorpromazine in a psychiatric hospital: one year's experience. *Marquette Med. Rev.*, 1955, **21**, 17–24.
6. Pascal, G. R., & Suttell, Barbara J. *The Bender Gestalt Test; Quantification and validity for adults*. New York: Grune & Stratton, 1951.
7. Rosen, A. Test-retest stability of MMPI scales of a psychiatric population. *J. consult. Psychol.*, 1953, **17**, 217–221.
8. Wilcox, P. H. The Gardner Behavior Chart. *Amer. J. Psychiat.*, 1942, **98**, 874–880.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Use of the Proverbs Test for Differentiating Schizophrenics from Normals[1]

Donald R. Gorham
Veterans Administration Center, Waco, Texas
Baylor University, Waco, Texas

In the Proverbs Test (10), the subject is asked to tell the meaning of a proverb. For normal subjects, even as young as fifth graders, this request readily stimulates what Goldstein refers to as the "abstract attitude." Indeed, the process of translating the concrete symbols of a proverb into generalized concepts is almost explicitly described in one of Goldstein's characterizations of abstraction:

> In the abstract attitude, we transgress the immediately given specific aspect or sense impression; we abstract from particular properties. We detach ourselves from the given impression, and the individual thing represents to us an accidental sample or representative of a category (7, pp. 18–19).

Benjamin (3) believed that the breakdown of verbal abstraction so commonly present in schizophrenia could be investigated by proverbs. Goldstein (8) claimed that the findings of the test for abstraction and concreteness which he and his associates had developed for investigating organic damage had relevance for research with schizophrenia and pointed out the need for more measures of abstraction.

The clinical Proverbs Test consists of three parallel forms containing 12 proverbs each, scaled in the order of difficulty. Samples are:

1. Rome was not built in a day.
2. When the cat's away the mice will play.
3. The wife is the key to the house.
4. Speech is the picture of the mind.
5. Better be happy than wise.
6. The more cost, the more honor.

An abstract (*Ab*) score on this form of the test is secured by rating the items according to sample scores for each proverb.

The multiple-choice Proverbs Test consists of 40 items with four answers each. It is scored by stencil. An abstract (*Ab*) score consists of the total number of correct responses. A concrete (*Co*) score consists of the number of items marked from 18 "concrete" answers. Samples of an easy and a difficult item are:

1. The sun shines upon all alike.
   a. It's the same sun everywhere.
   b. All are created equal. (*Ab* answer)
   c. The sun shines on everybody. (*Co* answer)
   d. People that do the same things are alike.
2. The good is the enemy of the best.
   a. Get the best of your enemy whenever you can.
   b. It is best to be good even if you fail.
   c. We should love our enemies.
   d. Most people are too easily satisfied. (*Ab* answer)

The Proverbs Test seems to possess useful qualities as an instrument for measuring impairment of conceptualization in schizophrenia. It gives a relatively pure measure of "verbal Comprehension" as evidenced by high correlations with tests heavily loaded with this factor (9, p. 7).

These considerations led the author to explore the possibilities of utilizing the Proverbs Test to investigate impairment in abstraction in schizophrenia and to differentiate schizophrenics from normals. Reliability studies indicated that the test was satisfactory when the subjects were hospitalized schizophrenic patients (10). As samples of schizophrenic populations were gathered, it became apparent that their Proverbs Test scores were generally lower than those of normals. Since test scores are so closely correlated with intelligence, however, this factor had to be controlled before it could be determined whether an actual difference existed.

---

[1] A paper presented at the annual meeting of the Southwestern Psychological Association, Dallas, Texas, March 23, 1956.

A pilot study was made in which 78 normal subjects were compared with 118 chronic schizophrenics. Analysis of covariance was used to hold intelligence constant and the Wechsler vocabulary subtest was used as the measure of intelligence. The analysis yielded an F of 14.30 which is significant at the .001 level of confidence. This finding seemed to justify a rigid test of the variation of scores on the Proverbs Test between schizophrenics and normals.

Method

From a sample of 332 Air Force basic airmen and 232 hospitalized schizophrenic veterans, 100 AF men were matched for sex, education, and intelligence with 100 chronic schizophrenics. The Word Knowledge Test of the Airman Classification Battery was used as the measure of intelligence. The pairs were matched exactly for stanine scores on the Word Knowledge Test. They were matched within two years for level of education; about two thirds matching exactly. The findings of this study are presented in Table 1.

The abstract (*Ab*) score and the concrete (*Co*) score on the multiple-choice Proverbs Test are secured by the use of scoring stencils. Comparable scores for the individual (clinical) form of the test are secured by individual scoring according to scoring samples for each proverb (9, 10). The table shows that the differences in the means of both the abstract and the concrete scores for both forms of the test are significant beyond the .001 level of confidence.

Since intelligence has been held constant by matching each normal with each patient on the Word Knowledge Test, the striking drop from normal to schizophrenic group on the abstract scores and the corresponding rise in the concrete scores indicate an impairment in abstraction resulting from the schizophrenic condition of the patients. In this connection it is interesting to note the negative correlation between the abstract and concrete scores on the Proverbs Test where independently derived. The coefficients of correlation between the *Co* of the multiple-choice test and the *Ab* score of the clinical test were found to be − .58 for 100 normal subjects and − .54 for 173 chronic schizophrenic subjects.

Cutting Points and Their Application

In order to establish cutting points and norms for use with the Proverbs Test the total available population of normals and schizophrenics was divided into intelligence groupings according to scores on the Word Knowledge Test. Such treatment of the data was necessitated by the correlation between

Table 1

Validity of the Proverbs Test for Discriminating Between Normal and Schizophrenic Men with Education and Intelligence Held Constant

(100 matched pairs: $N = 200$)

Test and score	Air Force enlistees		Chronic schizophrenics		Mean difference		
	M	*SD*	M	*SD*	D	t	p
Multiple-Choice Test							
Ab (Abstract) score	24.6	7.4	16.7	8.7	7.9	10.34	$\ll .001$
Co (Concrete) score	1.4	2.3	4.8	3.5	3.4	9.50	$\ll .001$
Clinical Test (Form I)							
Ab (Abstract) score	11.5	6.3	5.2	6.5	6.3	34.45	$\ll .001$
Co (Concrete) score	1.8	2.6	4.7	3.5	2.9	32.22	$\ll .001$

Table 2

A Comparison of Scores of Normals and Schizophrenics
Means and *SD*s by Intelligence Groupings *

Test and score	*Ab* (abstract) score		*N*	*Co* (concrete) score	
	Mean	*SD*		Mean	*SD*
Multiple-choice test					
High Intelligence (W.K.: 7, 8, 9)					
Normals	31.3	3.9	58	0.3	0.7
Schizophrenics	24.8	7.3	65	2.8	2.9
Middle Intelligence (W.K.: 4, 5, 6)					
Normals	22.8	5.4	136	1.6	2.4
Schizophrenics	15.3	6.5	79	5.1	3.6
Low Intelligence (W.K.: 1, 2, 3)					
Normals	16.0	4.9	58	3.9	3.0
Schizophrenics	9.6	3.9	65	6.1	2.9
Clinical test					
High Intelligence (W.K.: 7, 8, 9)					
Normals	19.2	2.6	58	0.2	0.6
Schizophrenics	12.6	6.4	65	1.9	5.9
Middle Intelligence (W.K.: 4, 5, 6)					
Normals	11.4	5.7	136	1.1	1.6
Schizophrenics	4.9	4.7	79	2.5	4.4
Low Intelligence (W.K.: 1, 2, 3)					
Normals	4.0	4.0	58	3.1	2.7
Schizophrenics	1.0	1.0	65	3.9	2.4

Note.—The normal population consisted of Air Force basic airmen from the Personnel and Training Research Center, Lackland A. F. Base. The schizophrenic population consisted of a representative sample of testable chronic hospitalized veterans from Waco Veterans Administration Center. These data are taken from norms in the clinical manual of the Proverbs Test (10) and are used with the permission of the publisher.

* Based on the Word Knowledge Test of the A. F. Airman Classification Battery. The W. K. scores 7–9 include the top 23%, 4–6 include the middle 54% and 1–3 include the bottom 23%.

intellectual level and performance on the Proverbs Test. The norms by intelligence groupings are presented in tables in the clinical manual of the Proverbs Test (10). In these tables the mean score and *SD* of normals and schizophrenics can be compared for each stanine score group of the Word Knowledge Test. Some of these data are presented in Table 2, which shows combined groupings of stanines 1–3, 4–6 and 7–9 on the Word Knowledge Test. These divisions are roughly equivalent to the upper quartile, middle half, and lower quartile in a distribution of normal subjects.

The data in Table 2 were used as the basis for establishing experimental cutting points for differentiating normals from schizophrenics. Actually, a considerable amount of trial and error entered into the process. It was found that at the upper and middle intelligence levels, the abstract and concrete scores of the multiple-choice test could be used most efficiently in combination. This was not true at the low intelligence level, where the concrete score had no usable differentiating value (see Table 2). The effectiveness of the cutting points when applied to a population of 352 normals and 232 chronic schizophrenics is shown in Table 3. The fact that a substantial portion of these cases was included in the population from which the cutting points were derived should be taken into consideration in judging the effectiveness of this application.

Table 3

Application and Cross Validation of Cutting Points for Multiple-Choice Test

	Original sample			Cross validation sample		
Group and cutoff scores	Normal	Schizophrenic	Total	Normal	Schizophrenic	Total
High Intelligence (W.K.: 7–9)						
$Ab \geq 26$ and $Co < 2$	86	22	108	48	6	54
$Ab < 26$ and/or $Co \geq 2$	19	43	62	9	21	30
	105	65	170	57	27	84
Middle Intelligence (W.K.: 4–6)						
$Ab \geq 15$ and $Co < 4$	112	23	135	24	8	32
$Ab < 15$ and/or $Co \geq 4$	29	56	85	4	22	26
	141	79	220	28	30	58
Low Intelligence (W.K.: 1–3)						
$Ab \geq 12$	70	18	88	16	10	26
$Ab < 12$	16	70	86	4	22	26
	86	88	174	20	32	52
Combined Groups						
Predicted Classification						
Normal	288	63	351	88	24	112
Schizophrenic	64	169	233	17	65	82
	352	232	584	105	89	194

The table shows that there were 64 false positives in the normal population sample and 63 false negatives in the schizophrenic population sample. The predictive efficiency of the application of the cutting points was: for normals, 81%, for schizophrenics, 73%, and for the total population of 584 cases, 77.5%.

In attempting to establish cutting points for the clinical test, it was found that the concrete scores could not be used successfully in combination with the abstract scores. The latter scores used alone, however, separated normals from schizophrenics with about the same degree of efficiency as the multiple-choice test scores. The application of the cutting points to the 100 matched pairs described earlier in this paper is shown in Table 4. The table shows that there were 17 false positives in the normal group, 28 false negatives in the schizophrenic group or a total of 45 misclassifications in the total sample of 200 cases. The efficiency of the clinical test for this sample is thus 77½%.

Cross Validation

The cutting points for use with the multiple-choice test were applied to a new population of normals and schizophrenics in order to cross validate the effectiveness of the measures. The new normal population consisted of 105 men and women in civilian and civil service employment at a variety of vocational levels. The new schizophrenic population consisted of 51 male veterans and 38 male and female chronic state hospital patients. The results of the cross-validation application of the cutting points for the multiple-choice test are given in Table 3.

The predictive efficiency of the cutting points applied to this new sample of 194 cases was 78%—almost identical with the original findings. The new sample contained men and women in both groups, confirming the normative data on the Proverbs Test which showed no sex differences (9). The inclusion of state hospital patients in the new sample indicates that the experimental cut-

Table 4

Application of Cutting Points for Individual Test *

Group and cutoff scores	Normal	Schizophrenic	Total
High Intelligence (W.K.: 7–9)			
Abstract score ≥ 14	24	8	32
Abstract score < 14	5	21	26
	29	29	58
Middle Intelligence (W.K.: 4–6)			
Abstract score ≥ 8	38	14	52
Abstract score < 8	10	34	45
	48	48	96
Low Intelligence (W.K.: 1–3)			
Abstract score ≥ 2	21	6	27
Abstract score < 2	2	17	19
	23	23	46
Combined Groups			
Predicted Classification:			
Normal	83	28	111
Schizophrenic	17	72	89
	100	100	200

* Population; 100 normals (Air Force basic airmen); 100 schizophrenics (chronic hospitalized veterans).

ting points have validity when applied to other than the male veteran schizophrenic patients on the basis of whose scores they were evolved.

Discussion

Several questions are raised in connection with the utilization of the Proverbs Test for differentiating schizophrenics from normals. The comparatively poor results of the multiple-choice test at the high intelligence level stand out. An examination of the actual cases in the schizophrenic group showed that 56% of the false negatives were paranoid schizophrenics, whereas only 34% of the true negatives fell in this diagnostic group. This finding points to the relative ineffectiveness of the multiple-choice test for screening out paranoid schizophrenics.

It is recognized, of course, that the efficiency of prediction is affected by the type of population to which the method is applied. A population containing roughly 50% normals and 50% schizophrenics does not actually exist. If applied to an unselected population containing only 1% schizophrenics, the predictive efficiency would be greatly reduced. On the other hand, if applied to a clinic population consisting entirely of people seeking assistance for mental and emotional problems, the predictive efficiency might well be enhanced.

When used with schizophrenics, the free-answer clinical test differs from the multiple-choice test in that it elicits psychotic material. In addition to concreteness in thinking which is associated with a breakdown of the abstracting function, the schizophrenic frequently "escapes" by a further regression characterized by "originality of performance" (1, p. 317). This is evidenced on the clinical test by responses characterized by autism, bizarreness, irrelevance, perseveration, and delusional preoccupation. Because of this free flow of schizophrenic material which is stimulated by this form of the Proverbs Test, diagnosis of schizophrenia is sometimes obvious upon clinical evaluation even though the objective scores fail to properly classify the subject.

It should be re-emphasized that the Word

Knowledge Test, which was used as the measure of intelligence in this investigation, is highly correlated with the Proverbs Test and gives a relatively pure measure of "verbal comprehension" (9, Table 3). Other testers will find the basic data for establishing intelligence divisions and cutting points suitable for their own purposes in the clinical manual of the Proverbs Test (10).

Summary and Conclusions

1. In a rigid statistical test to determine its effectiveness for discriminating between normal and schizophrenic subjects, both multiple-choice and individual clinical forms of the Proverbs Test yield mean scores whose differences are highly significant.

2. Cutting points for differentiating normals from schizophrenics are presented for three intelligence groupings which approximate the top quartile, middle half, and lowest quartile. The application of these cutting points correctly screens out approximately 80% of normal subjects and 75% of schizophrenic subjects. This level of efficiency was maintained on a cross-validation study based on 194 cases.

3. The effectiveness of the screening is consistent in the three intelligence levels for normal subjects. For schizophrenics the screening is least effective for the high intelligence grouping, probably due to the ineffectiveness of the technique when applied to chronic paranoid schizophrenics.

Received April 3, 1956.

References

1. Arieti, S. *Interpretation of schizophrenia.* New York: Brunner, 1955.
2. Baumgarten, F. A proverb for attitude measurement. *Personnel Psychol.*, 1952, **5**, 249–261.
3. Benjamin, J. D. A method for distinguishing and evaluating formal thinking disorders in schizophrenia. In J. S. Kasanin (Ed.), *Language and thought in schizophrenia.* Berkeley: Univer. of California Press, 1944. Pp. 65–90.
4. Bolles, M., & Goldstein, K. A study of the impairment of "abstract behavior" in schizophrenic patients. *Psychiat. Quart.*, 1938, **12**, 42–65.
5. Cohen, J. A comparative analysis of the factors underlying intelligence test performance of different neuropsychiatric groups. Unpublished doctor's dissertation, New York Univer., 1950. Univer. Microfilms, 1950, No. 1923.
6. Fruchter, B. The factorial content of the Airman Classification Battery. *USAF, Personnel Train. Res. Cent., Res. Bull.*, 1949, No. 49-1.
7. Goldstein, K. Methodological approach to the study of schizophrenic thought disorder. In J. S. Kasanin (Ed.), *Language and thought in schizophrenia.* Berkeley: Univer. of California Press, 1944. Pp. 17–40.
8. Goldstein, K. The significance of special mental tests for diagnosis and prognosis in schizophrenia. *Amer. J. Psychiat.*, 1939, **96**, 575–588.
9. Gorham, D. R. A proverbs test for clinical and experimental use. *Psychol. Rep. Monogr. Suppl.*, 1956, No. 1.
10. Gorham, D. R. *Proverbs Test. Clinical Forms I, II and III. Best Answer Form. Manual. Clinical Manual.* Louisville, Ky.: Psychological Test Specialists, 1956.
11. Hanfmann, E., & Kasanin, J. Conceptual thinking in schizophrenia. *Nerv. ment. Dis. Monogr.*, 1942, No. 67.
12. Hayakawa, S. I. *Language in action.* New York: Harcourt Brace, 1941.
13. King, W. R. Ability to abstract. In F. A. Mettler (Ed.), *Selective partial ablation of the frontal cortex: a correlative study of its effects on human psychotic subjects.* New York: Hoeber, 1949. Pp. 218–238.
14. Loeb, E. The function of proverbs in the intellectual development of primitive peoples. *Scientific Mon.*, 1942, **74**, 100–104.
15. Meadow, A., Greenblatt, M., & Solomon, H. C. "Looseness of association" and impairment in abstraction in schizophrenia. *J. nerv. ment. Dis.*, 1953, **118**, 27–35.
16. Pounders, C. J. A study of the Proverbs Test as a measure of the abstract level of concept formation. Unpublished master's thesis, Baylor Univer., 1956.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

The Ability of Schizophrenics to Modify Responses in an Interpersonal Situation[1]

M. David Diamond

Riverside Hospital, The Bronx, New York

The present paper discusses the results of an experiment designed to test the hypothesis that schizophrenics, as compared to nonschizophrenic emotionally disturbed individuals, are deficient in the ability to modify their responses in the direction of the responses of another person.

Mead defined rationality as the process whereby the individual "can take the attitudes of the others and control his action by these attitudes, and control their action through his own" (2, p. 339). Thus the "rational" person can modify his responses in the light of the responses of others.

Along similar lines of thought, Cameron states that, for the normal person:

> . . . the more effectively he is able to allow the attitudes and responses of others, which he predicts in symbolic role-taking, to influence his own reactions, the more competent he ought to be in social situations (1, p. 93).

Specifically the hypothesis tested was that schizophrenic individuals modify their responses less in accordance with the responses of another person when these responses are communicated in an interpersonal situation than do nonschizophrenic individuals. This hypothesis was tested at two levels of remoteness from immediate stimulation by having the subjects respond, first in the presence of another individual, and subsequently alone.

If an individual in an ambiguous situation modifies his responses in accordance with the responses of another individual, when he is in the presence of the other person, this would indicate that he can at least momentarily share the points of view of another person. If the individual modifies his responses in accordance with the responses of another individual, when the other individual is no longer physically present, this would indicate that his ability to accept the other person's points of view is not merely a transitory one, and that he is able to internalize the points of view of another. As a rough analogy we might, from a psychoanalytic point of view, consider the first task to be a measure of ego controls and functioning and the second one of superego controls and functioning.

Sherif's work with the autokinetic phenomenon suggested a technique for demonstrating the degree to which one's responses are modified by the responses of another, and the degree to which the responses of the other person are internalized. The autokinetic phenomenon consists of the perception of apparent movement of a stationary point of light, and is measured by having the subjects estimate the distance through which the point of light seems to move. Sherif found that when normal individuals are faced with the ambiguous autokinetic phenomenon, they form their own frame of reference and observe the point of light to move within a certain range. When individuals first face the autokinetic situation alone and then face it again in a group situation, they tend to change their estimates toward a norm and range which is established in the group. When subjects first face the autokinetic situation as members of a group, a norm and range are established which are peculiar to the group, and after-wards when the subjects face the same situation alone, they tend to adhere to the norm and range of the group. Sherif views the entire process as a prototype of the formation of a social norm, which then acts as a frame of reference for determining later reactions (3).

---

[1] This paper is a condensation of a part of a doctoral dissertation completed at New York University. The author wishes to acknowledge the help of Drs. L. Crafts, D. Wechsler, M. Deutsch, I. Katz, and P. Stout.

The autokinetic phenomenon seemed to be particularly appropriate for studying the effect of social interaction on behavior because it enables the investigator to isolate the variable of social influence from complex field influences such as the effects of learning and experience. Thus, in the present study an autokinetic experiment was used as a measure of the degree to which subjects modified their responses in the direction of the responses of others.

Subjects

The experimental subjects were thirty hospitalized male schizophrenics. Of the thirty schizophrenic subjects ten had been classified as schizophrenic reaction, paranoid type; nine as schizophrenic reaction, childhood type; seven as schizophrenic reaction, simple type; and four as schizophrenic reaction, chronic undifferentiated type.

A control group was sought which manifested a variety of types of psychopathology in order to eliminate the possibility that the hypothesized deficiencies in interpersonal skills were concomitant with emotional disturbance per se. Male drug addicts in a rehabilitation hospital were selected as control subjects since, as a group, they have been found to present a variety of emotional disturbances. No schizophrenic drug addicts were selected. Of the thirty drug addicts, eight were classified as inadequate personality, eight as schizoid personality; six as passive-aggressive personality; and one was classified in each of the following eight categories: adjustment reaction of adolescence, adjustment reaction of childhood, mixed psychoneurosis, neurotic character disorder, personality pattern disturbance, personality trait disturbance, psychoneurotic reaction, and sociopathic personality.

Table 1

Nonparametric Median Test for Initial Differences Between Pair-Members' Estimates of Light Movement When Judging Alone

Subjects	Below median	Above median	Total
Schizophrenic pairs	7	8	15
Drug addict pairs	8	7	15
Total	15	15	

Chi square = .13 $df = 1$ $p > .10$

Table 2

Nonparametric Median Test for Differences Between Pair-Members' Estimates of Light Movement When Judging Together

Subjects	Below median	Above median	Total
Schizophrenic pairs	4.5	10.5	15
Drug addict pairs	10.5	4.5	15
Total	15	15	

Chi square = 4.8 $df = 1$ $p < .02$

The control subjects were selected to be similar to the schizophrenics with respect to age, highest school grade completed, intelligence, race and socioeconomic background. The mean ages of the schizophrenic and drug addict groups were 19.3 and 18.5, respectively; the mean highest school grade completed, 9.6 and 9.9, respectively; the mean IQs 94.1 and 98.8, respectively. The two groups did not differ significantly with respect to any of these three characteristics. Both groups consisted of about 50 per cent Caucasian, and 25 per cent each of Puerto Rican and Negro individuals, primarily of lower socioeconomic background.

Procedure

The modification of responses in accordance with the responses of others was measured by means of an autokinetic experiment, in which the subject judged the apparent amount of movement of a stationary point of light in a darkened room, first when alone, then with another individual, and then finally alone again. Each subject made twenty estimates when alone. On the basis of these initial estimates when judging alone, subjects in each group were then assigned to pairs so that the initial differences between pair-members' estimates were approximately the same for subjects in both groups. The subject then made twenty judgments when with his pair-member and finally twenty more judgments when alone again.

The relative amount of convergence of pair-members' estimates when judging together provides a measure of the degree of modification of responses when in an interpersonal situation. The relative amount of convergence of estimates of pair-members when later judging alone provides a measure of the degree to which the "group norm" was introjected or internalized.

For each of the three parts of the experiment—individual judgments, judgments in pairs and subsequent individual judgments—the median of the twenty estimates was determined for each subject. The difference between the medians of pair-members' estimates for subjects in the two groups was then compared by means of the nonparametric median test for each part of the experiment. The nonparametric median test is essentially a chi-square test.

Results

Table 1 presents the results of the nonparametric median test which tests the hypothesis that the initial difference between pair-members is the same for the two groups. The resulting chi-square value of 0.13, which is not statistically significant, supports the null hypothesis that the two groups come from populations having the same median difference between pair members with respect to their assignment to pairs. Thus the pairs in the two groups can be considered similar with respect to initial differences between the medians of their estimates, before the experimental variable was introduced.

When judging with the pair members, subjects in both groups modified their responses in accordance with those of the pair members. However, the modifications made by the drug addicts were greater than those made by the schizophrenics.

Table 2 presents the results of the median test for the differences between median estimates of pair members when judging together. The obtained chi-square value of 4.8 is significant at the .02 level. Thus, the data support the hypothesis that schizophrenic individuals modify their responses in accordance with the responses of another person less than do nonschizophrenic drug addicts, when responding in the presence of the other person.

The results obtained with the drug addicts are similar to those obtained by Sherif when he performed an autokinetic experiment with pairs of normal subjects. When responding in a group after having first responded alone, the median difference between Sherif's pair-members' estimates was about 0.50 of an inch, which is the value obtained for the drug addict pairs (3).

When responding alone after having judged with the pair member, the drug addicts continued to modify their responses in accordance with those of the pair member as much as when they were in the presence of the pair member. The schizophrenics, on the other hand, tended to deviate more from their pair-members' estimates when no longer in the immediate presence of the pair member.

Table 3 presents the results of the nonparametric median test for the median difference between pair-members' estimates when judging alone after judging in pairs. The obtained chi-square value of 6.5, which is significant at the .01 level, supports the hypothesis that, after having participated in an actual interaction with another person, schizophrenic individuals modify their responses less in accordance with the responses of the other person than do nonschizophrenic drug addicts.

Table 3

Nonparametric Median Test for Differences Between Pair-Members' Estimates of Light Movement When Judging Alone After Judging Together

Subjects	Below median	Above median	Total
Schizophrenic pairs	4	11	15
Drug addict pairs	11	4	15
Total	15	15	

Chi square = 6.5 $df = 1$ $p < .01$

In this part of the experiment, also, the drug addict pairs resemble Sherif's normal pairs (3), in that the median difference between pair-members' estimates when judging alone after judging together was about 0.50 of an inch for both Sherif's group and the drug-addict group.

Discussion of Results

The results support the hypothesis that schizophrenic individuals modify their responses less in accordance with the responses of another person, when these responses are communicated in an interpersonal situation, than do nonschizophrenic individuals.

The subject's estimates of light movement when judging with another person, might be considered, from a psychoanalytic point of view, to reflect the adequacy of the subject's ego functioning. The person whose ego functions adequately would be expected, in the ambiguous autokinetic situation, to utilize the only frame of reference available besides his own—the responses of the other person—as a guide in making his own subsequent estimates. The results indicate that the schizophrenic has difficulty in utilizing such an external frame of reference.

The experimental evidence would suggest, further, that the schizophrenic does not adequately internalize the points of view of others. Since the superego is that aspect of the personality that is largely developed through a continuous process of introjection, these results suggest that, in the schizophrenic, not only ego functioning, but also superego functioning is affected.

Summary

A variation of the autokinetic experiment was employed to test the hypothesis that, as compared to nonschizophrenic emotionally disturbed individuals, schizophrenics are deficient in the ability to modify their responses in the direction of the responses of another person. Thirty hospitalized male schizophrenics and thirty male drug addicts judged the apparent amount of movement of a stationary point of light in a darkened room, first when alone, then with another individual from his group, and then when alone again.

The results supported the hypothesis. When estimating light movement in the presence of another person, the convergence of the schizophrenic pairs was significantly less than that of the nonschizophrenic drug-addict pairs. When responding alone, after having responded in the presence of the other person, the convergence of the schizophrenic pairs again was significantly less than that of the nonschizophrenic drug addict pairs.

Received February 21, 1956.

References

1. Cameron, N. *The psychology of behavior disorders*. Boston: Houghton Mifflin, 1947.
2. Mead, G. H. *Mind, self and society*. Chicago: Univer. of Chicago Press, 1934.
3. Sherif, M. A study of some social factors in perception. *Arch. Psychol. N. Y.*, 1935, No. 187.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Concept Learning in the Educationally Retarded Child of Normal Intelligence

Nissim M. Levy and Joseph M. Cuddy
Brown University

The report presented here is part of a larger project concerned with the development of a technique for predicting which children of normal intelligence, as measured by the usual methods, are likely to develop learning difficulties. The commonly-used measures in this area are estimates of intelligence, reading readiness, and social and physical development. Unfortunately these measures, except when they are fairly low, are relatively poor predictors.

Since most of the important tasks that children must master in school involve learning, a good instrument for estimating rate of learning may be of considerable value for evaluation purposes. This approach to the problem may be especially fruitful since rate of learning is very likely a complex function of not only intelligence, but also strength of motivation to learn, and reactions to failures during the acquisition process.

A careful examination of typical intelligence tests leads one to the fact that most of the material depends heavily upon the achievements of the individual. Of the WISC subtests, for example, none of them with the possible exception of the coding allow the subject an opportunity to improve performance during the assessment situation. This behavior sample is probably too small to be adequately used as a learning measure. However, it is interesting to note that Altus (1) reported consistently low scores on this subtest in records of 24 severe cases of educational retardation.

The present study was designed to answer the following question: Given a group of children of normal intelligence some of whom are underachievers in school, do underachievers differ significantly from normal achievers in ability to solve a concept-formation problem? A concept-formation problem was chosen for the first investigation on the assumption that it might provide the most sensitive measure of impairment in learning attributable to motivational and emotional difficulties.

Method and Procedure

The *Ss* were 23 pairs of fourth-grade elementary school children matched for age, sex, and, roughly, for socioeconomic status of parents.[1] The *Ss*' IQs ranged from 98 to 103 as measured by the California Test of Mental Maturity. One student of each pair was working up to grade placement while the other was behind from .5 to 2.5 years in achievement as measured by the Stanford Achievement Battery.

The learning task used for the study was a variation of the Oddity problem. Ten sets of Vigotsky Blocks were used. Each set consisted of three wooden blocks. In each set, the correct block was the odd one in terms of one of three possible dimensions. The three dimensions used were size, color, and shape.

Each *S* was told that his problem was to find out how to tell which was the correct block each time he was presented with a different set. If *S* chose the correct block, he received a chip. If he chose an incorrect block, *E* said, "That's not right," and the set of blocks was removed and another set presented. At the end of the game he received a penny for each five chips he had won.

---

[1] The authors wish to thank Dr. James Hanley, Superintendent of the Providence City Schools, and Miss Mary Lauth for their cooperation and assistance in making this study possible.

Five blocks of ten trials were given to each *S*, each block consisting of a random order of the ten different sets. If *S* chose correctly on 11 consecutive trials, testing was discontinued and *S* was credited with correct choices for the balance of the 50 trials. Testing was continued even when *S* spontaneously verbalized the correct principle before the end of the 50 trials. At the end, all *S*s were asked to verbalize how they solved the problem.

The measures of learning used were: (*a*) number of errors; (*b*) number of trials to criterion; and (*c*) verbalization of the Oddity principle.

Results

Errors. In 18 out of the 23 pairs of *S*s in the study, the normal achiever made less errors than his or her matched underachiever. Unless otherwise indicated all results reported were tested for significance by a two-tailed application of the binomial expansion. The probability of getting 18 out of 23 differences in one direction is less than the .01 level.

Trials to criterion. Eleven *S*s reached the criterion of 11 consecutively correct choices within 50 trials. Of these 11 *S*s, 10 were normal achievers while only one was an underachiever. This result yields a *p* value of less than .01.

Correct verbalization of the "Oddity" principle. Sixteen *S*s correctly verbalized the principle. Of these 16 *S*s, 14 were normal achievers while two were underachievers. This result yields a *p* value of less than .01.

Ten *S*s met both criteria. All 10 of these *S*s were normal achievers. Again this result is significant at well below the .01 level of confidence.

Using either criterion, 13 of the 23 pairs of *S*s had a difference in favor of the normal achiever, while no difference occurred in 10 pairs. Inspection of the data suggested the presence of a relationship between degree of discrepancy in achievement between the normal achiever and the underachiever of a given pair, and the likelihood of a difference occurring in either criterion in favor of the normal achiever. Therefore the difference score in achievement level for all of the pairs of *S*s was calculated. The over-all median of the 23 discrepancy scores was calculated and the data analyzed by means of the nonparametric median test. The contingency table arising from this analysis is presented in Table 1. Fisher's exact method for testing the significance of contingency tables with small frequencies was used and resulted in a *p* value of .01, confirming the suggestion of a relationship between the degree of underachievement and the likelihood of a difference in favor of the normal achiever.

It should be pointed out that two-tailed tests are conservative estimates of significance in this study since the implicit hypothesis being tested calls for a prediction of a directional difference. We would predict from both an empirical and theoretical basis that the underachievers would do less well.

Table 1

Median Test for Discrepancy Scores of Matched Subjects

Condition	Above median	Median or below
Normal achiever reached criterion; underachiever did not.	10	3
Neither *S* reached criterion.	2	8
Exact *p* value, .012		

Discussion

The results of the study are quite clear. Every measure of learning used demonstrated a significant difference in performance between our two groups. The finding that pairs of *S*s with higher discrepancy scores more often yielded a difference in criterion measures is further support for the hypothesis that underachievers learn more slowly.

Since 29 of the 46 *S*s failed to meet either criterion it is quite likely that the limit of 50 trials was too low. An examination of the verbalization of the *S*s who did not solve the problem indicated that many of them were at least on the way toward solution. When asked how they tried to solve the problem, some of them indicated partial solution by suggesting that it was the smaller one or larger one or the one that was a different shape. The *S*s were credited with the correct solution only if they indicated all three dimensions or simply said that the correct block was always different from the other two.

A difference in the rate of learning of a concept has been demonstrated in this study. The question arises as to whether this difference is restricted to the conceptual area or whether it may be a more generalized deficit. In an attempt to answer this question the same children are being used as *S*s in a number of other types of learning problems such as pursuit-rotor, stylus maze, paired associates, mirror drawing, discrimination, and problem boxes.

Whether we find underachievers demonstrating a general deficit in learning rate or whether it turns out to be primarily related to the verbal or conceptual areas the question will still remain whether or not we can detect these potential underachievers in preschool or the first grade.

Much of the thinking of clinicians concerned with the problem of the underachiever suggests that in many cases the familial interpersonal constellation and early history of the child play a considerable role in development of this condition. It would be reasonable to assume then that these children have, by the time they reach school age, already developed those patterns of behavior which will result in underachievement. This suggests that we can hope to find a learning-rate difference in children of normal intelligence by the time they reach school age which may be predictive of later underachievement in school.

It must be emphasized that what we are suggesting is not the use of learning tasks as a substitute for the standard intelligence scales. Hodges (4) following a lead of Hunt's (5) reported a study in which he estimated mental age from a variety of measures of performance in a double-alternation problem. The correlations which Hodges obtained varied between .22 and .74. It should be clear from his results that these two techniques are different.

Binet, it will be recalled, defined intelligence as the ability to profit from experience. Most intelligence tests use the same assumption that Binet used in order to justify themselves. This assumption is that all subjects measured by the particular instrument had the same opportunity for the experiences which are relevant to the material being used as the instrument. That this assumption is a weak one has adequately been pointed out before (2, 3). In part it is the weakness of this assumption which makes the use of such tests relatively inadequate for the prediction of academic success. Perhaps if Binet had worked on his problem after Thorndike, the tests which he would have developed might have been more along the lines of the work we are suggesting.

Summary

Twenty-three pairs of fourth-grade children matched for age, sex, and very roughly for socioeconomic status served as *S*s in a concept-learning task. The *S*s' IQs ranged from 98 to 103 as measured by the California Test of Mental Maturity. One *S* of each pair was working up to grade placement while the other was behind from .5 to 2.5 years in achievement as measured by the Stanford Achievement battery. All *S*s were given a maximum of 50 trials within which to reach the criterion of 11 consecutively-correct choices. A variation of the "Oddity" problem was used. At the end of testing, all *S*s were asked to verbalize how they attempted to solve the problem. The measures of learning were: (*a*) number of errors; (*b*) number of trials to criterion; and (*c*) correct verbalization of the oddity principle.

In 18 out of 23 pairs of *S*s, the normal achiever made fewer errors than his matched underachiever. Eleven *S*s reached the criterion of 11 consecutively correct choices. Sixteen *S*s correctly verbalized the principle. Fourteen were normal achievers while only two were underachievers. Ten *S*s met both criteria and all 10 of these *S*s were normal achievers. All of these results were statistically significant below the .01 level of confidence.

Implications of these results were discussed and the direction of further needed research was suggested.

Received July 26, 1956.
Early Publication.

References

1. Altus, G. T. A WISC profile for retarded readers. *J. consult. Psychol.*, 1956, **20**, 153–154.
2. Eells, K. Some implications for school practice of the Chicago studies of cultural bias in intelligence tests. *Harvard educ. Rev.*, 1953, **23**, 384–397.
3. Davis, A. Social-class influence upon learning. (The Inglis Lecture 1948) Cambridge, Mass.: Harvard Univer. Press, 1948.
4. Hodges, A. Double alternation: A measure of intelligence. *J. consult. Psychol.*, 1956, **20**, 59–62.
5. Nichols, P. C., & Hunt, J. McV. A case of partial bilateral frontal lobectomy. *Amer. J. Psychiat.*, 1940, **96**, 1063–1083.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

The TAT and Antisocial Behavior[1]

Kenneth Purcell [2]
Mental Hygiene Consultation Service, Fort Dix, New Jersey

Antisocial behavior may be viewed in terms of disequilibrium between an impulse system and a control system, the latter involving assessments of reality consequences as well as internalized standards serving as restrictions upon behavior. The present study is designed to examine certain disturbances of the impulse-control balance as they are manifested in TAT productions. Lindzey has referred to the "determination of the conditions under which inferences based on the projective material directly relate to overt behavior and the conditions for the reverse" (2, p. 18) as one of the major interpretive problems for the TAT. Studies examining the relationship between fantasy and overt behavior have yielded varied results (5, 9). Neither Murray (5) nor Sanford *et al.* (9) found any substantial direct relationship between the intensity of fantasy aggression and its overt expression. However, in explaining their respective results, each hypothesizes a factor of cultural prohibition serving to prevent the overt gratification of antisocial needs. In order to circumvent the complicating factor of middle-class inhibitions regarding aggressive behavior, Mussen and Naylor (6) studied a group of lower-class boys whose culture presumably encourages rather than inhibits overt aggression. They found strong support for a positive relationship between fantasy aggressive needs and overt aggressive behavior. Moreover, they found suggestive evidence that the amount of punishment anticipated relative to aggressive needs was an important determinant of overt aggression. With the exception of Mussen and Naylor's work, previous investigations attempting to correlate covert needs and overt behavior have tended to neglect the role of inhibiting or defensive forces in modifying the behavioral resultant of covert needs. Mussen and Naylor, however, made no effort to distinguish between external and internal controlling forces.

Hypotheses

The major hypotheses of the present study may be stated with reference to different facets of the impulse-control balance:

A. Impulse System

1. There is a direct relationship between degree of aggressive, antisocial behavior, and the number of aggressive thoughts manifested in TAT stories.

2. The form as well as the number of TAT aggressions is related to the extent of overt aggressive behavior. More aggressive individuals are likely to express TAT aggressions crudely and directly, relatively unhampered by repressive processes.

B. External Control System

3. Overt antisocial behavior varies inversely with the anticipated amount of fantasy punishment from an external source.

4. Among antisocial individuals themes of external punishment are more likely to serve as justifications for aggression than among nonantisocial individuals. This hypothesis is derived from Redl and Wineman's (8) detailed account of the "delinquent ego." By this is meant the ego whose every effort is bent toward obtaining guilt-free and anxiety-free enjoyment of delinquent impulsivity.

---

[1] The author wishes to express his appreciation to Richard Besnia, Carl Eisdorfer, Marvin Greenstein, Maurice Kouguell, and Dr. Philip Kubzansky who collected the data for this study and assisted in its statistical evaluation.

[2] Now at the University of Kentucky.

C. Internal Control System

5. Overt antisocial behavior varies inversely with the anticipated amount of fantasy punishment from an internal source.

6. Of the two types of punishment, internal and external, anticipated internal punishment is the more significant inhibitor of antisocial behavior.

Method

Fifty-seven male Army trainees between the ages of 17 and 26 with a mean age of 20 were used as subjects (*S*s) for this experiment. All were psychiatric referrals to the Mental Hygiene Clinic for whom psychological testing was requested. The only patients excluded were those suffering from organic brain damage or those diagnosed as psychotic.

Each *S* was seen individually by one of three clinical psychology specialists who administered cards 1, 3BM, 4, 6BM, 7BM, 8BM, 12M, 13MF, 14, and 18BM of the TAT. A fourth clinical psychology specialist and the author, neither of whom had administered any of the TATs, independently scored the TAT protocols for measures of aggressive impulses with respect to both quantity and quality, anticipated punishment from external agencies, and anticipated punishment from within.

The criteria used for identifying fantasy aggressions and punishments were based on those described by Mussen and Naylor (6) with certain modifications. A fantasy aggression score was derived for each *S* by counting the frequency of instances of all aggression such as the following: fighting, criminal assault, getting angry, criticizing, running away, resisting coercion, being negativistic, lying, cheating, stealing, dominating or restraining someone, rejecting someone. The occurrence of death, illness, or accident to parents or other loved objects in a story was also regarded as fantasy aggression. Since the present study is concerned only with extrapunitive aggression, such items as suicide, self-injury, and self-depreciation were not scored as fantasy aggression. The Pearson *r*, used as the reliability measure in every case, showed the correlation between independent scorings to be .91.

An external punishment score was obtained by summing the frequency of such themes as the following when they were directed toward the hero: assault, injury, threat, quarreling, deprivation of some privilege, object or comfort, domination, physical handicap such as blindness, etc., rejection. Scoring reliability was .81.

Internally based punishment included suicide, self-depreciation, and feelings of guilt, shame, or remorse. Since a broad definition of punishment encompasses "injury to a loved object," instances of death, illness, or accident to parents or other loved objects were recorded in this category of internal punishment. In instances in which the hero of the story was the direct agent of aggression, e.g., husband striking his wife, internal punishment was not scored except where there was a clear expression of remorse or shame. By contrast, accidental injury, death, or illness of a loved one always received a scoring of internal punishment unless the hero specifically denied anything resembling guilt feelings. Scoring reliability was .89.

Remoteness of fantasy aggression from overt behavior was rated on a seven-point scale with one representing the most direct form of fantasy expression. For each instance of fantasy aggression counted in the TAT story, a remoteness rating was made. The following considerations were kept in mind when evaluating remoteness: (*a*) object of aggression (person, society, animal, inanimate object), (*b*) time (present, past, future), (*c*) place (customary habitat, other countries, other planets), (*d*) level (behavior, wish, memory, daydream, nightdream, special states of consciousness such as intoxication, dissociation, drug addiction, insanity), (*e*) social context and instigator (hero instigates aggression, merely tags along with crowd, or even has nothing directly to do with aggression, e.g. death of someone due to disease or accident). The reliability for the number of one and two remoteness ratings scored per *S* was .96.

Measurement of aggressive antisocial behavior was based on the social history data routinely obtained by the psychiatric social work specialists. Such information included a record of *S*'s military delinquencies, e.g. AWOL, refusal to obey orders, assault, etc., as well as *S*'s report of any delinquent civilian behavior, e.g., truancies, gang fighting, stealing, peddling of drugs, etc. The military information was, of course, easily verifiable but the data pertaining to premilitary activities were not, except in certain instances in which records were requested from reformatories or other institutions mentioned by *S*. Several factors, however, indirectly testified to the substantial accuracy of *S*s' accounts of their civilian behavior. First, the civilian and military behavior data were almost invariably consistent with respect to delinquent characteristics. Nevertheless, it was important to use the premilitary history to get a long-term picture of *S*'s characteristic aggressive pattern, particularly for those individuals whose time in the military was relatively brief. Second, a subsequent check of psychiatric diagnosis revealed no case in which there was gross discrepancy between the clinical impression of the psychiatrist and *S*'s report of his aggressive behavior pattern. Third, on every occasion in which independent confirmation of civilian behavior was available (15 cases), *S*'s own account was corroborated.

Data pertinent to the variable of aggressive behavior were extracted from each *S*'s social history by the three clinical psychology specialists who had been responsible for testing. The author and one of these specialists rated the behavior extracts for aggressivity on a seven-point scale with one representing the least degree of aggressive behavior. The possibility of knowledge of test scores influencing behavior ratings or vice versa was negligible for two reasons. First, the specialist who served as a judge to check the reliability of the behavior ratings had administered only 25 per cent of the total number of TATs, and second, the behavior ratings were accomplished two months prior to the TAT scoring. Sample extracts of the extremes of the scale are presented below.

Rating of 1: Steady school attendance record through fifth term high school. After leaving school worked his way up to level of assistant pressman in print shop at $110 per week. He was drafted from a firm which had employed him since 1948. Married and living with wife and child for three years. No military or civilian delinquencies reported.

Rating of 7: Long history of truancies, delinquency and classroom misbehavior. In reform school four years. Prior to induction was confined two years eight months for armed robbery. While AWOL came in possession of drugs and jailed one month. Assaulted several people. Enlisted in Army on advice of parole officer to get reduction of sentence. History of eight AWOLs and four courts martial. Now facing a general court martial, five years in prison, and a dishonorable discharge because he wrecked stockade dispensary.

The reliability of these behavior ratings was .91. In order to assure large enough groups for comparison the seven-point scale was then consolidated into a three-point scale, combining ratings of one and two; three, four and five; and six and seven. Hereafter, Group I of 24 *S*s refers to the least antisocial group, Group II of 19 *S*s to the intermediate, and Group III of 14 *S*s to the most antisocial group.

Table 1

Product-Moment Intercorrelation of TAT Variables

	FA	Remoteness / FA	Int. Pun. / FA	Ext. Pun. / FA
Remoteness / FA	.53*			
Int. Pun. / FA	−.72	−.58		
Ext. Pun. / FA	−.45	−.13	.38	
Int. Pun. / Ext. Pun.	−.65	−.13	.89	.01

* An r of .36 is required for significance at the 1% level.

Results

Since the interpretation of results depends in some degree upon the relative independence of the five TAT variables their intercorrelations are presented in Table 1. Most of the relationships are statistically significant as seems logically required by the fact that complementary aspects of the impulse-control factor of personality are represented. However, in nearly all instances, the amount of variance in one variable that can be accounted for by the variance in another (as measured by r^2) is very low to moderate. Therefore, each of the variables may be said to contribute some unique information beyond their common meaning.

As shown in Table 2 the analysis of variance indicates over-all significant differences between the means of the three groups. This does not tell us, however, how well the three groups are separated. Therefore, before describing these results in detail, more precise information on the differences between the groups is presented in Table 3. Two-tailed tests of significance were used. It is apparent that in nearly all instances each group is significantly different from the other two.

The difference between the means of the three groups in number of fantasy aggression themes, as shown in Block A of Table 2, is highly significant beyond the value of $p =$

Table 2

Analysis of Variance for Three Groups Differing in Degree of Antisocial Behavior

TAT Variables	Mean		Variance*	F	p
A. Fantasy Aggression					
Group I	6.08	Between Groups	136.69	23.98	.001
Group II	8.05	Within Groups	5.70		
Group III	11.64				
B. $\frac{\text{Number of 1 and 2 Remoteness Ratings}}{\text{Fantasy Aggression}}$					
Group I	.10	Between Groups	.71	35.50	.001
Group II	.17	Within Groups	.02		
Group III	.49				
C. $\frac{\text{External Punishment}}{\text{Fantasy Aggression}}$					
Group I	.87	Between Groups	.33	3.66	.05
Group II	.70	Within Groups	.09		
Group III	.62				
D. $\frac{\text{Internal Punishment}}{\text{Fantasy Aggression}}$					
Group I	.79	Between Groups	2.88	20.57	.001
Group II	.32	Within Groups	.14		
Group III	.08				
E. $\frac{\text{Internal Punishment}}{\text{External Punishment}}$					
Group I	1.03	Between Groups	3.63	12.96	.001
Group II	.52	Within Groups	.28		
Group III	.15				

* Between Groups $df = 2$. Within Groups $df = 54$.

.001. Since several authors have suggested the importance of social-class sanction as a factor governing the overt expression of covert aggressive needs, this variable was explored within the limits of the data. The bulk of the antisocial group was of lower-class background whereas the nonantisocial group was predominantly middle class (χ^2 significant at 5% level).

To press the examination further, the *S*s in Groups I and III (least and most antisocial) were then separated according to social class. Using the *u* statistic described by Lord (3), the significance of the difference in number of aggressions for middle-class *S*s only was calculated. The same was then done for lower-class members of Groups I and III. Group III *S*s produced significantly more aggressive themes than Group I *S*s regardless of social class. Although the number of *S*s on whom these comparisons are based are relatively few (15 middle-class *S*s in Group I and 5 in Group III, 6 lower-class *S*s in Group I and 9 in Group III), the p values were .02 and .01. These several findings confirm the validity of Hypothesis 1 which concerns the relationship between aggression in fantasy and behavior.

To evaluate Hypothesis 2 the ratio of one and two remoteness ratings to total fantasy aggression score was compared for the three groups. The results, shown in Block B of Table 1, reveal the difference between the means to be significant well beyond $p = .001$. Not only does the antisocial individual respond with more aggressive themes, but the quality of his fantasy aggression is far more direct and undisguised than is true for the nonantisocial individual.

Block C of Table 1 compares the ratio of external punishment to fantasy aggression for the three groups. The significance of the difference between the means attains a value of $p = .05$. Thus an important force inhibiting the nonantisocial group appears to involve some sort of anticipated punitive action from without, giving support to Hypothesis 3.

Hypothesis 4 calls for a closer look at the fantasied external punishments. These were separated into three categories according to the following principles: (*a*) external punishment occurring subsequent to fantasy aggression and as a direct consequence of such aggression, e.g., "he was sent to jail for stealing"; (*b*) external punishment occurring prior to fantasy aggression and used as a justification for such aggression, e.g., "she was mean to him so he hit her"; (*c*) external punishment having neither of the above relationships to fantasy aggression, e.g., "she held him back because she was afraid he might get hurt." The total number of external punishments is larger for Group III than Group I with $p = .05$. It is clear, however, that TAT themes of external punishment are far more frequently utilized as excuses for aggression by antisocial members of Group III ($p = .001$), thus confirming Hypothesis 4. The mean number of themes classified in this manner was 3.64 for Group III, 1.25 for Group I. External punishment themes as a direct consequence of aggression are also more common among Group III *S*s although the *t* value of 1.89 falls just short of significance. Since the number of postaggression punishment themes is obviously dependent upon the number of fantasy aggressions, an attempt was made to compare retaliatory punishment themes on a proportional basis. The data showed that the mean ratio of retaliatory punishment themes to fantasy aggression is identical for Groups I and III so that it was not even necessary to calculate a significance statistic. Interestingly enough, the number of external punishment themes unrelated to fantasy aggression is larger among the nonantisocial members of Group I ($p = .01$).

Block D of Table 1 indicates that the differences between the group means, with respect to the quantity of guilt and shame themes proportional to aggression, is significant at the value $p = .001$. Since the antisocial *S*s produced significantly more aggressive themes than the nonantisocial *S*s, one may wonder whether this alone is not the crucial determinant in creating differential values of the ratio of internal punishment to fantasy aggression. If one were able to find individuals in Groups I and III with aggressive impulses of similar strength, then the effect of internal punishment expectations could be independently evaluated. Overlapping in the number of fantasy aggressions between *S*s of Groups I and III permitted comparison of six members from each group. They were the six highest and lowest in Groups I and III respectively in number of fantasy aggression themes. The *u* statistic shows no significant difference between the two groups with respect to fantasy aggression. However, the proportion of internal punishment to fantasy aggression is much larger for the nonantisocial group ($p = .001$). The mean ratio was .11 for the six Group III members and .54 for the six Group I members.

Table 3

t Ratios of Group Differences *

Groups compared	Fantasy aggression	Number of 1 & 2 Remoteness ratings / Fantasy aggression	External punishment / Fantasy aggression	Internal punishment / Fantasy aggression	Internal punishment / External punishment
Group I vs. Group II	3.18	1.75	1.77	3.36	2.83
Group II vs. Group III	4.17	5.33	1.09	6.00	4.07
Group I vs. Group III	6.04	7.84	2.27	4.44	4.40

* Minimum *t* ratios required for significance: 1.69 = <.10, 2.04 = <.05, 2.75 = <.01, 3.64 = <.001.

The relationship between socioeconomic class and antisocial rating raises a similar question regarding internal inhibition as was posed for fantasy aggression. Does the difference in number of internal punishment themes for antisocial and nonantisocial groups hold true irrespective of social class membership? The *u* statistic shows clearly that Group I *S*s produce significantly more internal punishment themes than Group III *S*s regardless of social class. The difference between the middle-class members of Groups I and III is significant beyond the .001 level and among lower-class members of the two groups the difference achieves significance at the .01 level.

As shown in Block E of Table 1, the ratio of internal to external punishment themes differs significantly for the three groups ($p = .001$). The much higher ratio for Group I *S*s suggests that internal punishment anticipations are a more potent deterrent to antisocial behavior than are external punishment anticipations, supporting Hypothesis 6.

Discriminative Power

Since a widely used diagnostic instrument was employed in this study it was deemed desirable to obtain some measure of diagnostic power or forecasting efficiency as distinguished from statistical significance. In order to evaluate discriminative power accurately it is necessary to make three comparisons of overlapping scores: Group I with Group II, Group II with Group III, and Group I with Group III. In each instance the median of the overlapping scores of the two populations being compared was selected as the cutoff point. Thus there were three different cutoff points for three separate comparisons. The percentage of correct classifications is calculated from the number of cases in each two-group comparison falling on the predicted side of the overlap median score. The data in Table 4 suggest that the most discriminating variables are the two internal punishment ratios. In clinical practice, however, it would seem appropriate to approach this classification problem by judging the interrelationship among the factors. For example, one of the members of Group I had a high fantasy aggression score but fell well within the nonantisocial group in all other scoring categories.

Table 4

Discriminative Power of Indicators Obtained from the TAT *

Variable	Groups I & II	Groups II & III	Groups I & III
Fantasy Aggression	68	67	76
Number of 1 & 2 Remoteness Scores / Fantasy Aggression	58	82	90
External Punishment / Fantasy Aggression	63	58	66
Internal Punishment / Fantasy Aggression	77	88	97
Internal Punishment / External Punishment	70	'82	92

* Numbers in table refer to percentage of cases correctly classified.

Discussion

Since the finding of a positive relationship between fantasy and aggressive behavior held true for *S*s of the middle as well as the lower class some attempt at explanation seems warranted in view of the contrary results of Sanford *et al.* (9) and Murray (5). The range in severity of aggressive behavior among *S*s in the present study seems far greater than was true for either of the previous studies and this, of course, made for an easier discrimination. Furthermore, it is the author's impression, also suggested by Redl (8) and Feshbach (1), that certain individuals who characteristically behave in impulsive, antisocial fashion not only find fantasy an unsatisfactory substitute for direct action but may actually be stimulated and provoked to act out by a succession of onrushing hostile images.

The association between lower-class membership and antisocial behavior is perhaps partially reflective of the freer rein given aggressive impulses in this group as suggested by previous authors. It is important to note, however, that when antisocial behavior does occur, whether it be among middle- or lower-class *S*s, these individuals manifest stronger aggressive impulses and/or a weaker internal control system than nonantisocial individuals. Thus the distinguishing feature is the individual impulse-control balance, although specific cultural groups may tend to foster varying types of impulse-control balance with respect to aggression.

The analysis of remoteness of aggression is related to the problem of symbolic representation of impulses and conflicts in projective data. This investigation demonstrated that antisocial *S*s were more direct in their expression of aggression than nonantisocial *S*s. It appears, then, that defensive inhibitions alter not only the sheer quantity of aggressive themes but also the way in which they are formulated. This is a common enough clinical observation but one which does not seem to have been given sufficient weight in experiments concerned with the problem of inferring behavioral tendencies from projective material.

Anticipation of punishment as a consequence of aggressive behavior was no more prominent among nonantisocial *S*s than antisocial *S*s. This is not to say that such expectation may not serve to deter antisocial action in many instances. However, the results do suggest that it is not one of the critical distinctions between the two groups. On the other hand, a significantly larger number of punishment themes unrelated to aggression either as retaliation or justification was characteristic of Group I *S*s. The content of themes classified in this manner suggests a greater awareness and acceptance of external direction and restraint. This finding may relate to the demonstration by Pastore (7) of the importance of perceived arbitrariness of the frustrating situation as a determinant of aggressive reaction.

As noted earlier in this paper there is a marked tendency for antisocial *S*s to justify their aggressive reactions by pointing to the harsh treatment imposed on them by others. If antisocial individuals can prove that the world is against them, then there is no need to recognize any obligations to others. The very fact that they so frequently find it necessary to resort to this projection technique argues that the delinquent ego is busily engaged in disposing of still existing remnants of guilt. One of the implications of this state of affairs is that any general policy of retaliation only reinforces this defensive maneuver and makes it easier for the delinquent to pursue his antisocial course on a guilt-free basis. The basic importance of internal controls in the inhibition of antisocial behavior is amply demonstrated by the data. Certainly the evasion of guilt, shame, or remorse is a far more vital mark of the delinquent than any lack of awareness concerning potential external punishment consequent to antisocial behavior.

Another pertinent topic involves the hypothesis that fantasy behavior has a compensatory function. Tomkins (10) assumes this hypothesis in stating that TAT aggressive fantasies are found more commonly among neurotics and normals than antisocial characters. The results of the present study suggest that the opposite is the empirical fact, given *S*s differing widely in antisocial behavior. However, this finding is not necessarily contradictory to the fantasy substitution hypothesis. One possibility is that the more potent and numerous aggressive impulses of the antisocial *S*s obscure any relative increment in fantasy aggression in the nonantisocial *S*s. An alternative suggestion mentioned earlier is that fantasy is not universally employed as a defensive or adaptive measure. Clinical experience tells us that individuals differ widely in the use of fantasy and it may simply be that many directly hostile individuals have not learned to resort to fantasy for substitute gratification.

Summary

This study has relevance to two problems. Its first interest was to examine certain personality variables, particularly in terms of the impulse-control balance, in order to establish their relationship to overt aggressive tendencies. The TAT was the instrument through which information was gathered about the nature of and interrelationships between aggressive impulses, anticipated internal punishment, and anticipated external punishment. The second major issue concerns the relationship between fantasy and reality behavior and the task of formulating from projective data alone the various cues which might permit differential predictions about behavior tendencies insofar as overt expression is concerned.

Fifty-seven young adult soldiers, the large majority of them basic trainees and all of them psychiatric referrals to the Mental Hygiene Clinic, served as *S*s. On the basis of social history data covering aggressive and antisocial behavioral manifestations in both civilian and military life, each *S* was rated as to overt antisocial tendencies. It was found that antisocial *S*s differ from nonantisocial *S*s in quantity and quality of fantasy aggressions, the former producing not only a larger number of fantasy aggressions but also finding it less necessary to obscure or minimize the hostile impulse. The superego variable as measured by anticipated internal punishments was found to be of far greater significance in inhibiting antisocial behavior than fear of retaliatory punishment which did not even differentiate between groups. An interesting finding dealt with the tendency of antisocial *S*s to utilize themes of external punishment as a justification for their fantasy aggressions. Some of the implications of these results were discussed. An effort was made to deal with the problem of socioeconomic class factors as they might relate to the data. Finally, emphasis was placed on the fact that, for behavioral prediction from projective material, the same attention must be given to repressive, inhibitory forces manifested in the data as is given to the impulse or drive system.

Received January 25, 1956.

References

1. Feshbach, S. The drive-reducing functions of fantasy behavior. *J. abnorm. soc. Psychol.*, 1955, **50**, 3–11.
2. Lindzey, G. Thematic Apperception Test; interpretive assumptions and related empirical evidence. *Psychol. Bull.*, 1952, **49**, 1–25.
3. Lord, E. The use of range in place of standard deviation in the *t*-test. *Biometrika*, 1947, **34**, 41–67.
4. McNemar, Q. *Psychological statistics.* New York: Wiley, 1949.
5. Murray, H. A. *Thematic Apperception Test manual.* Cambridge: Harvard Univer. Press, 1943.
6. Mussen, P. H., & Naylor, H. K. Relationships between overt and fantasy aggression. *J. abnorm. soc. Psychol.*, 1954, **49**, 235–240.
7. Pastore, M. The role of arbitrariness in the frustration-aggression hypothesis. *J. abnorm. soc. Psychol.*, 1952, **47**, 728–731.
8. Redl, F., & Wineman, D. *Children who hate.* Glencoe, Ill.: The Free Press, 1951.
9. Sanford, R. N., Adkins, M. M., Miller, R. B., Cobb, E. A., and others. Physique, personality, and scholarship: A cooperative study of school children. *Monogr. Soc. Res. Child Develpm.*, 1943, **8**, No. 1.
10. Tomkins, S. S. *The Thematic Apperception Test.* New York: Grune & Stratton, 1947.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Situational Influences on Rorschach Responses

Edith M. Henry and Julian B. Rotter
The Ohio State University

The widespread use of "projective" tests for purposes of diagnosis and prediction of behavior has stimulated extensive research investigating the validity of the underlying assumptions of these tests. One of the most significant areas in which research has raised questions is that of the effect of immediate influences on projective test productions. In social learning theory (14, 15) such influences are conceptualized as the effect of the psychological situation on behavior.

A growing body of literature demonstrates that many factors other than an individual's basic personality structure are operating in any test situation, and that these factors are important determiners of the types of responses to be given. That is, individuals adapt their behavior to the total environment in which they find themselves, and external stimuli as well as inner motivations and dynamics shape behavior at any given time. The present study is an attempt to show further the importance of situational factors and to predict the types of changes in response which will occur as a result of a particular type of change in the testing situation utilizing the Rorschach Test.

More specifically this study is concerned with the subject's (*S*'s) belief regarding the purpose of the test as it affects his test responses. A number of previous investigations (3, 4, 7, 9, 10, 11, 12, 16) have dealt with other kinds of situational influences on the Rorschach test, including such variables as the effect of immediate stress, the sex and personality of the examiner, suggestion, and the physical place of testing. There have also been a number of studies more closely related to the present one.

Fosberg (5, 6) found little consistent effect of asking *S*s to make the best possible and the worst impression without, however, clarifying how they should do this. Carp and Shavsin (2) however in a similar study found intra-individual differences although there were no statistical group differences. That is, *S*s changed their responses considerably but they did not do so in a consistent manner. Luchins (13) in investigating why *S*s did not give certain kinds of responses in the early part of the Rorschach test found that one of the more frequent reasons was the *S*'s fear of being labeled "psycho." Other reasons also implied that the *S* was responding to his interpretation of the purpose of the test. Similarly, Hutt, Gibby, Milton, and Pottharst (8) have shown that variations in instructions will produce significant variations in responses. Schachtel (17) has discussed in detail how the *S*'s implicit definition of the Rorschach test will affect responses but has not investigated these experimentally.

In the present study, the situational variable of the *S*'s definition of the test was manipulated and then investigated. It was felt that one of the most widespread views of the Rorschach, resulting from publicity of mass media, is that the Rorschach test is a measure of "sanity," that this test will uncover imbalances in a person's emotional life. In accord with the previous discussion, it was hypothesized that such a view of the test would alter an individual's responses in the direction of caution and the giving of only what most *S*s would consider to be "safe" responses. In order to test these assumptions, the experimental *S*s were given instructions about the test which included the suggestion that this test is designed to discover emotional disturbances of a serious nature in mental patients. An underlying assumption of the study was that, when *S*s respond in a particular way after being given a hypothesis about the test, these responses are comparable to those of another individual who holds the same hypothesis implicitly. The following hypotheses were formulated:

The experimental group will give significantly; (*a*) fewer responses in total (*R*) than will the control group; (*b*) more responses scored as form (*F*) than will the control group; (*c*) more responses of good form-level (*F* +) than will the control groups; (*d*) more large detail (*D*) responses than will the control group; (*e*) fewer whole (*W*) responses than the control group; (*f*) more popular (*P*) responses than will the control group; (*g*) more animal (*A*) responses than will the control group; (*h*) fewer responses of an aggressive and violent nature than will the control group.

In all these hypotheses the assumption is made that *S*s in the experimental situation will give, in the main, those responses which they consider to be "safe," inhibiting those which they consider suspect, those which would tend to place them in an unfavorable light. That is, they will give more responses which are conforming, which they expect to be similar to what most other people would see on the same blot or area. The present study differs from other Rorschach studies in that a situational variable is manipulated, and hypotheses are formulated about the changes in the determinants on a logical, "common-sense" basis rather than on the basis of the usual theoretical meaning of the various determinants although the two approaches coincided in some instances. The hypothesis about decrease in total number of responses for the experimental *S*s, for instance, was formulated because of the general tendency toward caution and inhibition. Similarly, an increase in pure form responses was hypothesized not because of its theoretical meaning as a measure of intellectual control but rather because form responses are often easier to see and those based on color, texture, or movement are frequently less clear and would be considered suspect. Large detail responses, again, were believed to correspond to the more obvious percepts; the preponderance of popular responses is in line with a "common-sense" attempt at caution and safety, at seeing things as most other people are likely to see them. The hypothesized increase in animal responses concurs with an increase in popular responses. The assumption that a large number of animal responses suggests intellectual immaturity was not employed. It was felt that a reduction in aggressive and violent types of responses would result from attempts by *S*s to make their responses as innocuous as possible, to inhibit any signs of lack of control over feelings.

Procedure

Subjects. Sixty female college students in elementary psychology were randomly allotted to two groups of thirty.

Procedure. When *S* entered the experimental room, she was asked the following initial questions:

> "Have you ever heard of the Rorschach test . . . the inkblot test?" If the response was affirmative, the following question was asked:
>
> "Tell me what you remember about it? . . . What do you think it does? What is it for?"
>
> Responses to these questions were recorded by *E* who then gave the main instructions. For the experimental group these were:
>
> "This is a series of cards with inkblots on them. People look at the cards and tell what the blots remind them of. For many years the test has been used in mental hospitals to study emotional disturbances and to discover how seriously disturbed patients were. Now we are making a college survey with it.
>
> "Now I want you to look at the blots and tell me what they could be, what they look like to you, or what they remind you of. People see all different things in the blots—there are no right or wrong answers."

The individuals in the control group, every second *S*, were also asked the questions about their previous acquaintance with the test, and then were given the standard Klopfer instructions:

> "This is a series of cards with inkblots on them. I want you to look at them and tell me what they could be, what they look like to you, or what they remind you of. People see all different things in the blots—there are no right or wrong answers."

Following this, the Rorschach was administered in the standard manner to all *S*s, and a fairly thorough inquiry followed.

After the test, the experimental *S*s were asked a number of general questions, designed to elicit information about the effectiveness of the instructions.

Scoring. All identification was removed from the records, and the protocols were scored according to Beck's (1) manual. An independent rater then scored twelve records, also using Beck's criteria. A short scoring manual was constructed for the presence of content involving aggression or violence. The categories used were as follows: explosions; fire; blood; people or animals battered, beaten, or demolished; fighting or killing; monsters involved in some kind of threatening action. A frequency count of these themes was then made by the experimenter and by the same independent rater.[1]

Statistical methods. To test reliability, a percentage of agreement was computed between the two scorers on twelve records. A Pearson product-moment correlation was used to test reliability of scoring for aggressive content on all sixty protocols.

To test the hypothesis dealing with total number of responses, a *t* test for differences between means was used. For all other hypotheses, two different techniques were used. The first technique involved converting all scores to percentages and then computing *t* tests for differences between the means of the percentages. The second technique was essentially a method of equalizing the scores to account for any differences in productivity (Total *R*) between the two groups.[2]

Results

Reliability of measures. Percentages of agreement were calculated between the scoring of *E* and an independent rate on twelve protocols. (On the remaining protocols, the *E*'s scoring alone was used with all identification first removed from the records.) Separate computations were performed for location, determinants, and content. The percentage of agreement for scoring of location was .95, for determinants, .82, and for content, .92. A second reliability check concerned the scoring for aggressive and violent content, scoring having been based on a brief manual constructed by *E*. Here, all sixty protocols were used. Pearson product-moment correlations were used here, with separate computations for the experimental and control groups. The correlation coefficient for the experimental group was .83, and for the control group, .66.

The results of the tests of the experimental hypotheses are given in Table 1.

a. *R* (total responses) lower in experimental than control group: The experimental group mean was 16.07, and the control group mean was 23.47. A *t* of 4.11 supports the hypothesis beyond the .001 level of confidence.

Table 1

Sigmas, Means, *t*s and Significance of Differences Between Means for the 30 Experimental and 30 Control Subjects

Variable	σ_e	Mean$_e$	σ_c	Mean$_c$	*t*	*p*
Responses	6.20	16.07	7.52	23.47	4.11	<.001
			From Equalized Scores			
Popular	1.85	.56	1.97	−.56	2.23	<.05
Form	2.43	.76	3.56	−.75	1.88	<.10
Form-Level	1.17	.66	2.49	−.67	2.62	<.02
Aggressive	1.94	−.48	1.80	−.12	1.28	<.20
Animal	2.14	.73	2.86	−.74	2.25	<.05
Whole	3.48	.15	5.04	.87	.63	<.60
Detail	3.91	.23	3.97	.22	.02	<.90
			From Percentages			
Popular	.12	.32	.04	.26	2.29	<.02
Form	.19	.72	.13	.61	2.50	<.02
Form-Level	.08	.87	.08	.80	2.70	<.01
Aggressive	.11	.07	.10	.10	1.08	<.20
Animal	.14	.53	.12	.43	2.93	<.01
Whole	.23	.45	.46	.44	.92	<.40
Detail	.24	.50	.70	.49	.40	<.70

---

[1] The authors would like gratefully to acknowledge the assistance of Mr. Walter Mischel in this respect.

[2] We are grateful to Dr. Robert Wherry who suggested a method for equalizing the effects of an unequal number of responses in the two groups. To equalize scores the following steps were taken for each variable except *R*: (*a*) The total number of responses for both experimental and control groups was computed; (*b*) The total number of responses on any particular variable, such as *F*, were computed, and this number was divided by the total *R* to give a predicted score on the variable under consideration for the group as a whole; (*c*) To arrive at an individual predicted score, this total predicted score was multiplied by each individual's total *R*; (*d*) This predicted score was then subtracted from each individual's actual score on the variable being considered. This score served as an indication of where each individual's score fell in relation to the score that would be predicted from total group performance. These scores were then used in the computation of *t* tests of significance between the means of the two groups.

b. More *F* (form) responses in experimental than control group: Using equalized *R* scores, means of .76 and − .75 were obtained for the experimental and control groups, respectively. A *t* of 1.88 was obtained, which yields a *p* value between the .05 and .10 levels and tends to substantiate the hypothesis. Using percentage scores, the mean value for the experimental group was .72 and .61 for the control group. A *t* of 2.5 supports the hypothesis between the .02 and .01 levels of confidence.

c. More *F*+ (good form level) responses in experimental than control group: With equalized scores, means were .66 and − .67 for experimental and control groups, respectively. A *t* of 2.62 supports the hypothesis between the .02 and .01 levels. A *t* calculated from percentages yielded a value of 2.70, significant beyond the .01 level in the predicted direction. Mean values were .86 and .80 for the experimental group and the control group, respectively.

d. More *D* (large detail) responses in experimental than control group: With equalized scores, means of .23 and .22 were obtained for the experimental and control groups. A *t* of .02 and, with percentages, a *t* of .34 failed to give any substantiation to this hypothesis. Means of percentage scores were, respectively, .50 and .49.

e. Fewer *W* (whole) responses in experimental than control group: Using equalized scores, a mean of .15 was obtained for the experimental group and of .87 for the control group. A *t* of .63 with equalized scores and .92 with percentage scores did not substantiate this hypothesis. Means with percentage scores were .45 and .44 respectively.

f. More *P* (popular) responses in experimental than control group: Experimental group mean, with equalized scores, was .56; control group mean was − .56. A *t* of 2.23 with equalized scores lent support to this hypothesis between the .02 and .05 level. With percentages, means were .32 and .26, respectively. A *t* of 2.29 supported the hypothesis between the .025 and .01 levels.

g. More *A* (animal) responses in experimental than control group: Using equalized *R* scores, means were .72 and − .74, respectively. A *t* of 2.25 supports this hypothesis between the .02 and .01 levels. With percentages, means were .53 and .43 for experimental and control groups, respectively. With *t* calculated from percentage scores, significance was between the .01 and .001 levels.

h. Fewer aggressive responses in experimental than control group: With equalized scores, means were − .48 and − .12, respectively. A *t* of 1.28 fell between the .10 and .20 levels of confidence. Using percentage scores, means were .07 and .10, respectively. A *t* of 1.08 also fell between the .10 and .20 levels, and, therefore, did not significantly substantiate this hypothesis but was in the right direction.

Discussion

The results of this study, in general, seem to substantiate the hypotheses advanced. In regard to aggressive responses, the total number produced in the group was too few to obtain significance in the results and the assumption that the more conforming responses were frequently more *D* than *W* was not borne out. The other hypotheses regarding the number of responses, the number of form responses, form-level, popular responses, and animal responses, were all supported to a highly significant degree.

The real significance of this study does not lie in its demonstration of the effect of a different type of instructions but rather in the implications that subjective hypotheses about what the test is testing will have a significant effect on the *S*'s responses. As applied to the Rorschach test some specific information is presented in this study as to what that effect may be when the *S* has the hypothesis or belief that the test is essentially a test of insanity, mental disturbance, psychopathology, etc. Such a belief is probably fairly commonly held as a result of the widespread popularization of the Rorschach through magazine articles, Sunday supplements, movies, radio, television, and also through direct communication. The results support strongly Schachtel's assertion that it is extremely important for an examiner to find out what these implicit beliefs regarding the test are, and to evaluate test results in the light of them.

It is of some interest that the general questioning of the *Ss* following the test administration did not reveal too much in the way of awareness that the *Ss* were trying to appear in a good light, not mentally disturbed, etc. Some *Ss* after considerable pushing would finally agree that they tried to avoid mentioning anything that might sound peculiar, but in general, more prediction could probably be obtained by assuming certain effects likely to occur as a result of the *S*'s hypotheses about the test purposes than could be obtained by questioning the *S* regarding his motives in taking the test.

The study also tends to support, at least by implication, the general notion that conceptualizing the *S*'s behavior (including test responses) in a test situation as an attempt to reach some goal to give a particular kind of impression or to defend himself against some threat can be a useful predictive method from a common-sense point of view.

The results also tend to support the now growing body of literature which emphasizes the need to interpret psychological test results in light of a host of situational variables, no matter how indirect the test, rather than limiting analysis to the test protocols or profiles obtained from several tests or subtests.

Summary

This study measured the influence of one aspect of the psychological situation, knowledge of test purpose, on responses to the Rorschach test. The *Ss* were sixty female students in elementary psychology. Experimental *Ss* were told that the Rorschach is used to discover serious emotional disturbances, then given standard Klopfer instructions; control *Ss* received only Klopfer instructions. Hypotheses were that experimental *Ss* would give fewer total, aggressive, and whole responses; more pure form, good form-level, popular, animal, and large detail responses. The specific hypotheses all involved the general hypothesis that experimental *Ss* would be cautious and conforming. Results were analyzed by two methods for every variable except total responses. By one, raw scores were converted into percentages, and *t* tests computed. The second involved a technique of equalizing scores to eliminate any influence of reduced number of total responses in one group. With total responses, *t*'s were computed on differences between means. Results showed fewer responses in experimental group (beyond the 1% level), more good form-level, popular, and animal responses (between 5% and 1% levels), more form responses (between 10% and 5% levels), fewer aggressive responses (between 20% and 10% levels). Location did not vary significantly between groups.

The results tend to support the general hypotheses that: (*a*) a knowledge of the *S*'s beliefs regarding the purposes of the test should be an important aspect of the interpretation of the test, (*b*) a "common-sense" analysis of the *S*'s approach to the whole testing situation would be a fruitful part of the test interpretation, and (*c*) situational determinants significantly influence test responses.

Received January 16, 1956.

References

1. Beck, S. *Rorschach's test.* Vol. 1. *Elementary principles.* New York: Grune & Stratton, 1944.
2. Carp, A. L., & Shavzin, A. R. The susceptibility to falsification of the Rorschach psychodiagnostic technique. *J. consult. Psychol.,* 1950, **14**, 230–233.
3. Coffin, T. E. Some conditions of suggestion and suggestibility. *Psychol. Monogr.,* 1941, **53**, No. 4 (Whole No. 241).
4. Eichler, R. M. Experimental stress and alleged Rorschach indices of anxiety. *J. abnorm. soc. Psychol.,* 1951, **46**, 344–355.
5. Fosberg, I. A. Rorschach reactions under varied instructions. *Rorschach Res. Exch.,* 1938, **3**, 12–31.
6. Fosberg, I. A. How do subjects fake results on the Rorschach test? *Rorschach Res. Exch.,* 1943, **7**, 119–121.
7. Gibby, R. C., Miller, D. R., & Walker, E. L. The examiner's influence on the Rorschach protocol. *J. consult. Psychol.,* 1953, **17**, 425–428.
8. Hutt, M. L., Gibby, R., Milton, E. O., & Pottharst, K. The effect of varied experimental sets upon Rorschach test performance. *J. proj. Tech.,* 1950, **14**, 181–187.
9. Kimble, G. A. Social influence on Rorschach records. *J. abnorm. soc. Psychol.,* 1945, **40**, 89–93.
10. Klatskin, E. H. An analysis of the effect of the

to respond in the usual manner to Cards III and VII of the Rorschach. The sample was then divided at the median into High (2 or more) and Low (less than 2) *M* groups. It was determined that there was no significant difference between the two groups in the number of *S*s who responded with an overt expression of affect ($\chi^2 = .12$, $p = .70$). Thus the record was successful in inducing an impulse to laugh, and the expression of that impulse, regardless of the amount of *M* produced by the *S*s.

Sixty-eight additional *S*s from the same general population were introduced to the same experimental procedure with one crucial modification. Prior to hearing the record, *S*s were instructed *not* to smile or laugh. This sample was similarly divided at the median into high *M* and low *M* groups. The two groups were compared on their success or failure in controlling the expression of the experimentally induced affect.

It is to be noted that the Rorschach was administered *following* the laugh record in order to rule out any bias in the judgment of affective response. Thus the experimenter had no way of knowing whether a subject was going to fall in the high *M* or low *M* group at the time of making the judgment about the presence or absence of an affective response.

Rorschach protocols were scored independently by two judges without knowledge of the *S*s performance on the inhibition task. In the few cases of scoring disagreement the judgment of a third psychologist served as the deciding factor.

Results

Results were in the predicted direction. Of the 34 *S*s in the high *M* group, 20 were able to listen to the entire record without any observable display of feeling. Only 11 of the 34 *S*s in the low *M* group were similarly successful (Table 1).

Table 1

Affective Control of High and Low *M* Groups

Group	Successful Control	Failure to Control
High *M* (2 or more)	20	14
Low *M* (less than 2)	11	23

Since the direction was hypothesized, a one-tailed test of significance was used. The obtained chi square of 3.79 is significant at the .03 level. Analysis of the data in terms of the total number of responses (*R*) reveals that the obtained relationship is not dependent upon *R* ($\chi^2 = 1.27$, $p > .20$). Tetrachoric correlation between *M* and affective control was .42 ($p < .01$).

Discussion

These results support the hypothesis of a significant positive relationship between *M* and the voluntary inhibition of expression of an experimentally induced impulse to laugh. Although these data cannot be generalized to the control of other affects, one could hypothesize that further experimentation with other affects would reveal similar relationships.

It is of interest to speculate about the processes involved in controlling affective expression. Is there any fundamental difference in process between the inhibition of the *expression* of an affect and the control of the affect itself? The two may be equivalent only from the point of view of the outside observer. Controlling affective expression appears to have both motor and cognitive components. A person can keep from expressing an impulse to laugh, for example, by controlling certain facial muscles. The affect can remain even though its overt expression is controlled by motor means. Some of the subjects reported that they felt very much like laughing, but restrained themselves by clenching their teeth and compressing their lips.

The other way of controlling the expression of affect is through the modification of the affective tone itself by cognitive processes which alter the autonomic reaction pattern. By way of illustration, one of the subjects reported that she kept from laughing by thinking of her mother dying. She was thereby able to change the affective tone through cognitive means. The affect need not be changed to the extent of substituting an opposite one, as thoughts could be used in a variety of ways merely to neutralize the existing feeling. We have already cited experiments which have demonstrated that the abilities to inhibit motorically and cognitively are each correlated with *M*. If voluntary control of the impulse to express an affect is viewed in terms of its motoric and cognitive components, the present experiment may simply be reflecting these relationships in a more complex setting.

As suggested above, there may be two basic styles of inhibiting affective expression—one primarily motoric, and the other principally cognitive. We might speculate further and invoke a concept of motor-cognitive vicariousness. This would point up the possibility that motor inhibition tends to lead to the production of thoughts or fantasies which in turn serve to modify the affective tone. Observations and reports from the *S*s suggest that their initial adaptation to the task was through motor control. It appeared that this was later supplemented by cognitive methods at different stages according to the individual. This progression from motor to cognitive control can be thought of as a recapitulation of the developmental process, in that the ability to use cognitive processes to control feelings might be expected to appear at a later stage of personality development. Observations of young children suggest that control largely through motor means is gradually supplemented by the acquisition of a variety of cognitive techniques which can modify the affective state. When these mechanisms are refined in the normal adult, it is more difficult for the observer to detect the affective state than it is in the emotionally labile child. We would expect different behavorial and symptomatic manifestations in adults with these two different styles of control in accordance with the differences in their chronic reaction pattern. Thus we might anticipate different symptom complexes in individuals who are chronically exerting motor control in contrast to others whose control is exercised largely through thinking. It would therefore be important to contrast individuals who are found to utilize predominantly one or the other of these methods from the point of view of relating their style of control to developmental levels and to the complicated problem of symptom choice in psychopathology.

Summary

Previous experiments have demonstrated motor and cognitive inhibition to be behavioral correlates of Rorschach Human Movement responses. The present study extends the relationship of *M* and inhibition ability further to include the control of affective impulses as well. An impulse to laugh was induced, and *S*s instructed to refrain from any affective expression. In support of the hypothesis, more high *M* than low *M* subjects were successful at inhibiting any overt expression of the experimentally induced affect.

Received February 21, 1956.

References

✓ 1. Levine, M., & Meltzoff, J. Cognitive inhibition and Rorschach human movement responses. *J. consult. Psychol.*, 1956, **20**, 119–122.

✓ 2. Meltzoff, J., & Levine, M. The relationship between motor and cognitive inhibition. *J. consult. Psychol.*, 1954, **18**, 355–358.

✓ 3. Meltzoff, J., Singer, J. L., & Korchin, S. J. Motor inhibition and Rorschach movement responses. *J. Pers.*, 1953, **21**, 400–410.

✓ 4. Singer, J. L., & Herman, J. Motor and fantasy correlates of Rorschach human movement responses. *J. consult. Psychol.*, 1954, **18**, 325–331.

5. Singer, J. L., Meltzoff, J., & Goldman, G. D. Rorschach movement responses following motor inhibition and hyperactivity. *J. consult. Psychol.*, 1952, **16**, 359–364.

6. Singer, J. L., & Spohn, H. E. Some behavioral correlates of Rorschach's experience type. *J. consult. Psychol.*, 1954, **18**, 1–9.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

The Use of the Kaleidoscope as an Adjunct to the Rorschach

Emanuel Starer
VA Hospital, Northport, N. Y.

One of the problems encountered in the psychological testing of many chronic psychotic patients is that they frequently produce only ten or fewer responses on the Rorschach. As a result, the examiner is hampered in his attempt to describe the underlying personality structure in detail.

The writer has employed the kaleidoscope with 20 chronic patients, with the instructions to report whatever was seen. After the patient had finished associating to a given stimulus, he was instructed to shift the kaleidoscope which would automatically introduce another stimulus. In this manner an unusually large number of stimuli could be introduced in a relatively short space of time. Approximately 15 to 20 minutes were allowed for each patient. The associations to the kaleidoscope could be analyzed in a manner similar to content analysis on the Rorschach.

Most of the patients produced two to three times the number of responses obtained on the Rorschach. Indeed, one patient who refused the Rorschach volunteered to associate to the kaleidoscopic stimuli and produced 30 responses which were helpful in evaluating some of his areas of conflict.

Examples of some of the responses obtained with the kaleidoscope include those frequently obtained from the Rorschach such as: bugs, butterflies, bats, insects, plants, flowers, and designs, usually geometric in nature.

Responses containing a color component included: a bloody rag, two big red blotches, face of a cat with red eyes and red nose, and I see a discolored ball, sort of pale and brown.

Some of the responses containing human, animal, or inanimate movement included: man's face, sort of stupid looking with sharp nose, statue of the Lord, He's bowing His head, face of a cat who is ready to eat you up, crabs clawing an insect, a spider making up his mind whether to poke at me or not, three bolts of lightning hitting an object from three different directions.

Finally, some examples of highly personalized associations included: pair of ice tongs like the ones I used when a boy, type of Pennsylvania-Dutch design, a hex sign, insignia of the devil, fundus of the uterus, mop of hair of a child, pancake hat for young lady, and motor neurone.

The kaleidoscope technique is recommended only when there is a paucity of associations on the Rorschach, and so can be considered only an adjunct. Much valuable information is lost since only a content analysis can be made. Perhaps the most objectionable to many investigators is that the stimuli are not constant and so can not be reduplicated. In spite of these and other objections, this examiner has found the adjunctive use of the kaleidoscope in personality evaluation to be worthy of further investigation and use in practical clinical testing situations, particularly with chronic psychotic patients.

Received February 29, 1956.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

The Use of Human Figure Drawings in the Diagnosis of Organic Pathology

Marvin Reznikoff and Donald Tomblen
The Institute of Living

Clinical psychologists are often asked to aid in the differential diagnosis of organic pathology. One of the widely used techniques in psychodiagnostics is the human figure-drawing test. Experience has suggested that this relatively unstandardized method may prove to be useful in attempting to evaluate the presence or absence of organic difficulties, but there are relatively few published studies in this area. Michal-Smith (2) reported the results of the H-T-P technique, comparing boys whose EEGs were abnormal with boys who were found to be normal with reference to their EEGs. Only one sign yielded a significant difference, namely, "line quality."

Machover (1) and Vernier (3) both listed indicators which they mentioned as "often seen" in the organic case. For Machover, these included: disproportionately large head; empty, large, poorly proportioned figures; weakly synthesized drawings; and simple, heavy line quality. Vernier suggested "petal" or scribbled fingers; center-of-page placement; shrunken arms and legs; erasures in arm and hand area; and eyes drawn as slits. In agreement with Machover, she also listed largeness of head and whole figure; emptiness, or lack of detail; weak synthesis; and poor proportions. However, she described the line quality as sketchy and/or broken, rather than heavy and simple.

The present study was undertaken to investigate: (*a*) the degree to which these various suggested indicators are actually more characteristic of the drawings of organic patients than they are of the drawings of other psychiatric groups, and (*b*) the extent to which these indicators may reliably be employed in the differential diagnosis of organicity.

Procedure

Sample. Three groups of 25 patients each were selected from the files on the basis of final diagnoses, which in this institution follow the criteria and classifications described in the *Diagnostic and Statistical Manual of Mental Disorders* prepared by the American Psychiatric Association. In one group, all patients had been given organic diagnoses, specifically "chronic brain syndrome." Another group of patients were diagnosed as schizophrenic, and the last group as having neurotic disorders. Organic findings were entirely absent in the schizophrenic and neurotic groups. All patients were at least average in intelligence and between 19 and 55 years of age, with a mean age of 35.4 years. All patients had drawn human figures during routine psychological testing. The patients had not been tested by either author, nor was any information about the personal history, diagnosis, or results of other psychological tests available to the authors when they scored the drawings.

Scoring method. A list of 17 indicators was compiled, based upon those suggested in the literature plus several others derived from the personal experience of the senior author. Scoring criteria for each indicator were delineated. In the instance of No. 15, "Enlarged figure size," actual measurement of the drawing length was recorded. For all other indicators, a simple plus-minus system was used. If an indicator was clearly present, it was scored plus; otherwise, it received a minus score. The indicators were:

Table 1

Analysis of Variance of Drawing Size Measurements of Organic, Schizophrenic, and Neurotic Patients

Source of variation	*df*	Mean square	*F*
Between groups	2	6.24	1.15
Within groups	72	5.44	
Total	74		

1. Emptiness of figure (lack of details)
2. Weak synthesis (parts loosely joined)
3. Parts misplaced (noticeably shifted)
4. Disproportionately large head
5. Shrunken arms and legs (thinner and shorter)
6. Other parts distorted (inappropriate size and/or shape)
7. "Petal" or scribbled fingers
8. Eyes drawn as slits
9. Center-of-page placement
10. Erasures in arm and hand area
11. Erasures elsewhere (two or more in single drawing)
12. Heavy, simple line quality
13. Sketchy line quality
14. Broken line quality
15. Enlarged figure size
16. Unlifelike pose
17. Erasures with only minimal improvement

The drawings of all 75 subjects were randomly arranged with all identifying data removed. Each of the authors independently scored every drawing for every indicator. Two indicators proved to be unscorable in practice and therefore were eliminated. These were: No. 16, "Unlifelike pose," and No. 17, "Erasures with only minimal improvement," both original with the authors. It was impossible to arrive at a satisfactory, objective definition of unlifelike pose, and it was likewise extremely difficult to determine the degree of improvement after erasures without actually observing the drawing process itself. On the 15 retained indicators scored for every drawing, there was 91 per cent interscorer agreement. Minor differences of scoring were readily resolved in conference, still independent of any knowledge of the subject's diagnostic category.

Table 2

Drawing Indicators Showing Significant Differences Between Groups of Organic and Neurotic Patients of Twenty-five Each

Indicator	Frequency of occurrence		Chi square	*p*
	Organic	Neurotic		
Weak synthesis	15	1	15.53	.001
Parts misplaced	15	1	15.53	.001
Shrunken arms and legs	14	2	11.12	.001
Parts other than head and extremities distorted	15	3	10.50	.01
Petal or scribbled fingers	13	4	5.70	.02
Emptiness of figures	15	5	6.75	.01

Treatment of data. Indicator No. 15, "Enlarged figure size," was tested separately by means of an analysis of variance of figure measurements. All other indicators (Nos. 1–14) were studied together. A tabulation was made of the frequency of occurrence of each indicator among the 25 patients in each diagnostic group. Maximum frequency would thus be 25. Each group was then compared with the other two groups by means of chi square.

Analysis of variance was again utilized to test whether or not the patients in the organic group made scores significantly higher than patients in the other groups. For this particular analysis, the contradictory hypotheses concerning line quality were eliminated, and a score was obtained for each patient consisting of the number of indicators present out of a possible total of 11 (Nos. 1–11). These scores were converted to percentages, and an angular transformation was done in order to stabilize the variance. The analysis-of-variance design was then applied to the transformed scores of the 75 patients distributed into the three diagnostic groups.

Results and Discussion

From the results of the analysis of variance cited in Table 1, which yielded nonsignificant results, it is evident that the organic patient does not characteristically draw larger figures than other patients.

Tables 2 and 3 show that 5 of the 14 indicators are indeed more characteristic of organic drawings than of drawings of either schizophrenics or neurotics. An additional indicator, emptiness of figures, showed a significant difference between organic and neurotic productions, but not between organic and schizophrenic drawings. The remaining indicators occurred with approximately equal frequency in all groups, with any discrepancies attributable to chance factors. When schizophrenics were compared with neurotics, none of the indicators differentiated between the groups to a statistically significant degree.

It may be noted here that Vernier's sketchy or broken line indicator was present more often in the organic group than was Machover's simple, heavy line. However, neither of these indicators appeared to a greater extent in the organic group than they did in the other groups.

Table 3

Drawing Indicators Showing Significant Differences Between Groups of Organic and Schizophrenic Patients of Twenty-five Each

Indicator	Frequency of occurrence		Chi square	*p*
	Organic	Schizophrenic		
Weak synthesis	15	5	6.75	.01
Parts misplaced	15	5	6.75	.01
Shrunken arms and legs	14	0	16.77	.001
Parts other than head and extremities distorted	15	6	5.26	.05
Petal or scribbled fingers	13	1	12.00	.001

Results of a final analysis are reported in Table 4 and indicate the highly significant degree to which organic patients surpass both schizophrenics and neurotics in the number of indicators present in their individual drawings.

Since the analysis-of-variance design applied to the transformed scores demonstrated a disparity among the three groups significant at the .001 level, the groups were next considered two at a time. The *t* test was applied using the mean square value of the within-group source of variation (error term) as the pooled variance. Results of the comparisons are shown in Table 5. It can be seen that the scores of the organic patients differed from both those of the neurotics and schizophrenics at the .001 level of confidence, while the difference between scores of the schizophrenic and neurotic patients was not statistically significant.

Table 4

Analysis of Variance of Transformed Drawing Scores of Organic, Schizophrenic, and Neurotic Patients

Source of variation	*df*	Mean square	*F*
Between groups	2	1612.19	13.65*
Within groups	72	118.15	
Total	74		

* Significant at the .001 level of probability.

Although these indicators distinguish between groups, it remains to be determined how effectively they can be utilized in individual differential diagnosis. Table 6 reveals that as many as 7 out of 11 indicators must be present in order to establish with certainty that the correct diagnosis is an organic one, rather than neurosis or schizophrenia. With such stringent requirements, one would miss nearly three-quarters of all organic cases! Reducing the number of indicators to 5 would still result in 40 per cent false negatives and also in obtaining 4 per cent and 16 per cent false positives from among the neurotic and schizophrenic groups, respectively. Since these predictions are based on the same cases as were used for the analysis of the indicators, it is likely that a greater number of false diagnoses would occur on cross validation with a new sample. The evidence clearly points to the limitations of a naive reliance on a purely "sign" approach to the

Table 5

Separate Group Comparisons of Drawing Scores of Organic, Schizophrenic, and Neurotic Groups

Comparison	*t*	*p*
Organics with neurotics	4.66	.001
Organics with schizophrenics	4.16	.001
Neurotics with schizophrenics	0.49	not significant

Table 6

Cumulative Distribution of Drawing Scores of Organic, Schizophrenic, and Neurotic Patients

Drawing score*	Organics		Schizophrenics		Neurotics	
	Number of patients	% of patients	Number of patients	% of patients	Number of patients	% of patients
1 or more	24	96	24	96	24	96
2 or more	24	96	20	80	19	76
3 or more	21	84	15	60	14	56
4 or more	19	76	7	28	6	24
5 or more	15	60	4	16	1	4
6 or more	12	48	1	4	1	4
7 or more	7	28	0	0	0	0
8 or more	4	16	0	0	0	0
9 or more	1	4	0	0	0	0

* Drawing score equals number of indicators recorded as present out of maximum of eleven.

diagnostic task. Nevertheless, the sophisticated clinician may well utilize the more discriminating indicators as a basis for further investigating the possibility of organic pathology in a patient.

Summary

This study attempted to evaluate a number of drawing indicators cited in the literature as characteristic of organic patients. It was found that five of these indicators were more prevalent among a group of organic patients than among schizophrenic or neurotic groups. These were: weak synthesis; parts misplaced; shrunken arms and legs; parts, other than head and extremities, distorted; and petal or scribbled fingers. The remaining ten indicators did not differentiate between the three groups.

When the number of indicators present was scored for the patient's individual drawings it was found that the organics had significantly higher scores, but the groups overlapped to a considerable degree, limiting the usefulness of this approach alone for individual differential diagnosis.

Received February 24, 1956.

References

1. Machover, Karen. *Personality projection in the drawing of the human figure.* Springfield, Ill.: Charles C Thomas, 1949.
2. Michal-Smith, H. The identification of pathological cerebral function through the H-T-P technique. *J. clin. Psychol.*, 1953, **9**, 293–295.
3. Vernier, Claire M. *Projective test productions: I. Projective drawings.* New York: Grune & Stratton, 1952.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Certain Formal Characteristics of Human Figure Drawings by Institutionalized Aged and by Normal Children[1]

Martin Lakin

University of Illinois, College of Medicine
Orthodox Jewish Home for the Aged, Drexel Home for the Aged
Chicago, Illinois

Drawings of the human figure have been fairly widely used and studied as a psychological diagnostic and personality assessment technique. They have been used as bases for inferences about intellectual or maturity level, conflict areas, and self concept or "body image." The present study is an investigation of the extent to which figure drawings reflect self concepts or body images which may be assumed to be characteristic of disparate levels of age and social status. The specific question under consideration is whether figure drawing is a technique which is sensitive to *changes* in body image occurring from childhood to old age.

That changes in body image might be anticipated in individuals who experience the physiological vicissitudes associated with aging is suggested by Schilder; "When we build up a persistent and coherent image of the body we build it up out of our emotional mood which is based upon biological tendencies . . . the changes in the biological unit reflect themselves in the more plastic image of the body" (4).

Several studies have found that the drawings of institutionalized aged persons differ from the drawings of certain groups of normal adults. According to one recent study, as compared with the productions of university graduate students, those of the aged were characterized by incompleteness, lack of proportionality, and lack of integration (2). Machover (3) has observed that senile individuals draw small, puny figures, use relatively little area of the paper for the drawing and center the figure poorly upon it.

These observations are interesting in view of current theory with respect to the technique. In clinical work, gross departures from "normative" performances in terms of such formal variables as area, size, and centeredness are frequently interpreted as suggestive evidence of deviant intrapersonal states. Area utilized, for example, is considered to reflect self-assessment in relation to the environment. Height is similarly interpreted as an indication of feelings of relative self-significance. Centeredness is often considered to be affected by the subject's general security level and by his orientation toward reality.

Gross differences in self conceptualization and, presumably, in body image would be anticipated in a comparative study of the institutionalized aged and normal children close to the threshold of adolescence. In one way their life situations are somewhat similar: both groups are explicitly dependent upon protecting and nurturant environmental forces for gratification of primary needs. However, as contrasted with the vigor, growth, increasing self-awareness, and recognition of potential among the children, the psychological decrements which accompany the physiological changes in the aged are a matter of popular concern. Loss of self-esteem, lessened vitality, and the sense of impending further decline seem to underlie the frequently observed ex-

---

[1] Thanks are due Drs. Marion Wieman and Alan K. Rosenwald of the University of Illinois, College of Medicine for their advice and cooperation in this study.

Table 1

Identifying Data for Subjects

Mean	Age	Education (yrs. of formal schooling)	N Males	N Females
Institutionalized aged subjects ($N = 24$)	78.5	3*	13	11
Normal children ($N = 25$)	8.7	3	10	15

* Estimated because of discrepancies in education of aged who were all foreign born.

pressions of futility and hopelessness among the institutionalized aged. The question is whether these differences are reflected in the figure drawings of the two groups.

Such formal variables as area, height, and relative centeredness would appear to be excellent points of comparison between the two contrasting groups. They are elements which are relatively free from the influences of previous education and artistic training. They are much less subject to these influences and to conscious control than such features as integration of the figure, facial expression, and proportionality. Also, with respect to these latter elements, the influence of peripheral, sensorimotor impairments would have to be taken into account. In the present study of formal characteristics peripheral impairment would seem to be of minimal significance. If the formal characteristics, area, height, and centeredness do reflect self-conceptual or body-image dispositions, then it would be hypothesized that differential performance in terms of them would distinguish between the drawings of institutionalized aged and those of normal children.

Method

Twenty-five third grade children from a public school and twenty-four elderly residents of a home for the aged were the subjects. All were administered the Figure Drawing Test (3) and were provided with similar materials, medium soft pencils and plain $8\frac{1}{2} \times 11$ sheets of paper. The aged subjects were given the test in individual examination sessions. The children were tested as a group. Instructions were limited to the simple directions: "Draw a person; you may make it any way you wish, but draw a whole person, not just the face."

Table 1 describes the personal characteristics of the groups. The children were a typical class of third graders. Those markedly high or low in previously administered Stanford Binet scores (above IQ 110 or below IQ 90) were not included in the present sample. The institutionalized aged were also a fairly "typical" group, in terms of the home's standards. All were ambulatory and in "good" contact. Markedly deteriorated individuals and those considered to be grossly impaired intellectually as evaluated on the basis of the Bender Gestalt Test and psychiatric interview, were not included in the study.

Comparison of the aforementioned formal characteristics of the drawings were undertaken as follows:

Area. The area of each drawing was precisely measured by means of a geographer's planimeter. The instrument yields results in terms of square inches. The areas of the two groups, treated as scores, were then compared. Because of the somewhat skewed distribution of the data the nonparametric median test was employed.

Figure height. The distance from figure top to figure bottom was measured with an ordinary ruler. Heights to the nearest inch were treated as scores. The scores of the groups were then compared by means of the median test.

Centeredness. A grid of concentric circles, $\frac{1}{4}$ inch apart, was constructed. Each circle was numbered and regarded as a unit. By placing the grid-template over the sheet with the drawing, it was possible to measure the distance from the center of sheet to the center of gravity (the midpoint between the figure's extremities) of the figure. Measured distances from the center in units were termed scores and were compared by means of the median test.

Results

Table 2 shows the differences in the performances of the two groups of subjects.

On the measure of area the group of institutionalized aged and that of normal children had medians of 3.0 square inches and 6.8 square inches respectively. On the second measure, that of distance from the center in $\frac{1}{4}$ inch units, the medians were 11.0 and 6.0.

Medians of the group with respect to the measure of height were 4.0 inches and 6.0 inches respectively.

To provide some indication of the variability the number exceeding the combined median is given for both groups. For the measure of area the numbers are 6 and 18 respectively. For distance from center they are 17 and 8, and for height they are 6 and 19.

Discussion and Conclusions

The results confirm the hypothesis that formal characteristics of the human figure drawings differentiate two groups widely differing in age and social status, and assumed to be widely differing in self concept and body image.

The direction of the difference, the more constricted, shorter, and less adequately centered productions of the institutionalized aged, as compared with those of the children is congruent with the assumed differences in self-attitudes among these groups. What is significant, in terms of the present study, is that the feelings of self-devaluation and a "shrinking" body image should find graphic expression in the execution of formal features of the drawings. On the other hand, the possibility that certain organic factors in the aged result in constrictions of the figure is counterindicated by nonconstricted Bender-Gestalt drawings among this group.

The findings seem to give additional evidence of the validity of the assumption that formal aspects of figure drawings are related to the central variables of self conceptualization and body image. In light of the clinical assumption about departures from "normative" performance it is interesting to note that the children seem to perform more similarly to normal adults than do the aged subjects. Levy (1) reported an average figure drawing height of 7.0 inches for normal adults. The median figures for the aged and child subjects of the present study are 4.0 inches and 6.0 inches respectively.

The findings suggest the application of similar methodology in further investigation in this area. Comparison of drawings by institutionalized aged and by noninstitutionalized aged persons might be undertaken to study the question of whether changes in self concept and body image are due to psychological processes attendant upon aging per se or whether they are related to such factors as relative independence of functioning and the effects of permanent custodial-protective care.

Table 2

Differences in Area Size, Height, and Distance from Center of Figure Drawings of Institutionalized Aged Persons and Normal Children

	Group		
	Institutionalized aged	Normal children	Median test
Area (sq. inches)			
Median	3.0	6.8	
Number exceeding combined median	6	18	10.8**
Distance from center (¼" units)			
Median	11.0	6.0	
Number exceeding combined median	17	8	7.4*
Height (to nearest inch)			
Median	4.0	6.0	
Number exceeding combined median	6	19	12.7**

* Significant at the .01 level.
** Significant at the .001 level.

Summary

On the basis of reported observations and theory regarding human figure drawings and their relationships with self concept and body image it was considered that such formal characteristics of the drawings as area, height, and centeredness would reflect aspects of change in self concept and body image. The relative freedom of these variables from background influences and from the influences of peripheral impairments was suggested. It was hypothesized that differences in self conceptualization between institutionalized aged and preadolescent children would be reflected in differences in these formal aspects of the drawings. The drawings of 24 institutionalized elderly persons were compared with those of 25 third-grade school children. The re-

IDK—I Don't Know. When the patient says "I don't know," not in answer to a direct question, but as in resignation, disgust, or despair. Examples:

1. "So that's the way it happened. I don't know . . . (IDK) things were all mixed up."
2. ". . . and the pain has been there since then. (Pause) I don't know (IDK)."

SI—Sighing or Deep Breath.

L—Laughing. Any kind of laugh or chuckle.

V—Voice Change. Change in volume, becoming louder or lower.

?—Questioning the Interviewer. When the patient asks the interviewer to repeat what he said, not for elaboration but because he was not heard or understood. Examples:

1. P: "What?" (?)
2. P: "Pardon me?" (?)

BL—Blocking. When there is groping for the proper expression, indicating unusual hesitation. Examples:

1. "So she . . uh . . well . . . uh . . (BL) told me I shouldn't come back."
2. "So I . . so I (R) . . (snaps fingers) (BL) I (R) told her I wouldn't anyway."

To estimate reliability of scoring these cues, two raters trained together by reviewing and discussing the scoring of eight 10-minute samples of interviews. Then they separately scored cues on 10 interviews and their percentage of agreement on each cue was computed by dividing the number of agreements by the total of agreements plus disagreements. Agreement means both raters note the same cue in the same spot on a line of speech. Disagreement occurs when either rater notes a cue and the other calls it some other cue or leaves that spot unmarked.

Table 1

Percentage of Agreement of Two Raters on Each Cue

Cue	Agree/agree plus disagree	% Agreement
UN	82/134	61
BR	5/26	19
INT	24/72	33
R	143/213	66
ST	16/17	94
IDK	9/19	47
SI	14/14	100
L	22/22	100
V	41/41	100
?	4/4	100
BL	2/2	100

Ten-minute samples of the remaining 29 typescripts of interviews were then cue-counted and a Cue Score, expressed as a percentage, was computed for each interview by dividing the total number of times the cue was scored in the interview by the total number of typed lines of patient speech. To analyze the interrelationships of the cues, a contingency coefficient between each distribution of cue scores was then computed. The intercorrelations indicated the groups of cues which we called Cue Scale I and Cue Scale II.

Group cue scores for Cue Scale I and Cue Scale II were obtained by converting the cue score distributions over all interviews to percentile scores, then for each interview adding the percentile cue scores within each scale. Because of its negative correlation with V and SI, the L distribution was made negative before combining it with them.

Results

Table 1 presents the raters' agreement on scoring each cue. The perfect rater agreement on SI, L, V, ?, and BL results from the fact that they are clearly marked on the typescript by the transcribing secretary. UN, R, and especially ST have respectable per cent agreements, but it was obvious that if BR, INT, and IDK were to be used, they would have to be grouped with the more reliable cues.

Table 2 presents the intercorrelations of the cue scores. The two cues ? and BL were

Table 2

Intercorrelations of Cues
Contingency Coefficients Corrected for Class Index
(39 Interviews)

Cue	BR	INT	R	ST	IDK	SI	L	V
UN	.42	.20	−.03	.20	.56	−.20	−.13	.13
BR	—	.03	.13	.20	.03	.13	−.28	.13
INT	—	—	.36	.56	.60	.03	.13	−.28
R	—	—	—	.63	.13	.28	−.13	−.20
ST	—	—	—	—	.20	.24	−.05	.11
IDK	—	—	—	—	—	.28	−.28	−.03
SI	—	—	—	—	—	—	−.28	.69
L	—	—	—	—	—	—	—	−.28

omitted from consideration because they occurred rarely in the data and provided distributions with very little spread.

By inspection of Table 2, the cues were grouped into patterns containing the highest possible intercorrelations. The following groupings were prominent: BR-UN-IDK-INT-R-ST and L-V-SI. These groups we call respectively Cue Scales I and II. With the exception of IDK, the cues of Scale I are all indicators of speech disruption. Scale II is composed of cues related to nonverbal voice characteristics. The reliability of rating of the cues grouped in Scale I is 65 per cent and of those grouped in Scale II is 100 per cent. Cue Scales I and II have a Kendall's rank correlation (Tau) of $-.03$ ($p = .81$). They are unrelated and used independently as anxiety measures.

Discussion

In the author's study referred to earlier (1), the Cue Scales and three other indices of anxiety (skin conductance curves, clinicians' global ratings, and patients' self-reports) were used to measure patient anxiety in reaction to structured or ambiguous interviewing technique. The Cue Scales showed significant relationships to degree of ambiguity, just as did the other measures. The anxiety measures showed no significant correlation with each other, except for the correlations (.32 and .26) of the two Cue Scales with anxiety ratings of transcripts by clinical judges. This correlation indicates only that both sets of speech cues (uncorrelated with each other) enter into a judge's estimate of anxiety when reading transcripts. An analysis of the findings of the study supports the hypothesis that anxiety is not a unidimensional phenomenon but can be evidenced by activation of various physiological or behavioral systems in different persons. Thus a person who shows speech disruption in a situation of uncertainty need not show concomitant GSR changes. Is GSR a better indicator of anxiety than speech disruption? At this point it seems safe to say only that they are both methods by which we can identify anxiety in order to learn more about it. The speech disturbances measured by the Cue Scales are clearly related to situational anxiety and not to anxiety as a personality trait, for the patient groups showed significant speech differences in reaction to the varied interviewer technique although they were randomly selected and equated for diagnosis and several other pertinent variables.

The Speech Scales represent speech variables indicative of "emotionally involved" and disorganized verbal behavior. Some of the variables seem to be direct instances of disorganization (e.g., incomplete sentence or breaking in of new thoughts) and others as defenses against or attempts to control disorganization (e.g., repetition or modification of words and phrases). The lack of correlation (Tau = .01 and .07) of the two scales with the patient's own report of anxiety leads us to feel that the cues of disorganization, involvement, and defense, measured by the scale, are anxiety indicators not subject to the person's awareness. Further research should be done to explore other speech cues of anxiety, to evaluate the consistency of appearance of the cues under varying conditions of anxiety, and to investigate the relationship of the cues to individual speech habits. The use of the scales as tools to measure anxiety related to such therapeutic phenomena as interpretation, reassurance, or transference, as well as for evaluation of therapeutic progress should be explored.

Summary

The derivation of two measures of transient or situational anxiety as revealed in speech has been described. The two uncorrelated measures are based upon a count of speech cues reflecting disorganization, emotional involvement, and defensiveness. The use of the scales in a study of anxiety in psychological interviews was described, and some implications of the scales as anxiety measures were discussed.

Received March 12, 1956.

References

1. Dibner, A. S. The relationship between ambiguity and anxiety in a clinical interview. Unpublished doctor's dissertation, Univer. of Michigan, 1953.
2. Dollard, J., & Mowrer, O. H. A method of measuring tension in written documents. *J. abnorm. soc. Psychol.,* 1947, **42**, 3–32.
3. Kauffman, P. E., & Raimy, V. C. Two methods of assessing therapeutic progress. *J. abnorm. soc. Psychol.,* 1949, **44**, 379–385.
4. Lasswell, H. D. Verbal references and physiological changes during the psychoanalytic interview: a preliminary communication. *Psychoanal. Rev.,* 1935, **22**, 10–24.
5. Sanford, F. H. Speech and personality. *Psychol. Bull.,* 1942, **39**, 811–845.
6. Zimmerman, J. Modification of the DRQ as a measure of progress in counseling. Unpublished master's thesis, Univer. of Chicago, 1950.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Test Anxiety and Intelligence Test Performance

I. Zweibelson

Milford, Connecticut, Public Schools

Relatively few persons have escaped the worry which is usually associated with the group tests which are used extensively in our culture. It is also true that there have been relatively few studies of the influence of fear, worry, and anxiety on group test performance. Studies (3, 7) of the worries of elementary school children have revealed that they worried most about test failure. It has also been shown (8) that college students with high anxiety performed noticeably more poorly on group intelligence tests than students with low anxiety.

The Problem

The present study [1] concerns the relation between children's test worry or anxiety and their scores on group tests of mental ability which differed in content, format, and method of administration. The following question was posed: does performance in mental ability test situation x, y, or z differ in relation to the relative amount of test anxiety indicated by the subject? Table 1 briefly summarizes the important differences of the three mental ability tests used in the study: the Otis-Alpha, the Davis-Eells, and the Otis-Beta. An index of pupil test anxiety was secured by the investigator who read a list of 43 questions [2] to 12 classes of fifth grade elementary school children. Each child responded on an answer sheet by circling a *"Yes"* or *"No"* for each question. The questions dealt specifically with fears, worries, and feelings in test and testlike situations.

Hypotheses

1. Given three tests of mental ability, the Davis-Eells, the Otis-Alpha, and the Otis-Beta, performance on the Davis-Eells and the Otis-Alpha tests should be less related to test anxiety than in the case of the Otis-Beta.
2. In the case of the Otis-Beta and the Davis-Eells, the "High-Anxious" should do better on the mental ability test taken second than on the one taken first, regardless of the test, i.e., the order in which the tests are taken will have an effect on the scores of the "High Anxious."

Design

Five of nine elementary schools in Milford, Connecticut, were selected as schools which served town areas thought to be representative of pupils in all of the town schools. All of the fourth grade children in the selected schools were given three mental ability tests (1, 5, 6) and the Test Anxiety Questionnaire. The Otis-Alpha and the Questionnaire were administered in the ninth month of the fourth grade school year. The Otis-Beta and the Davis-Eells were administered during the first month of the fifth grade school year. Half of the pupils were given the Otis-Beta first and half the Davis-Eells first—the total N was 258. There was approximately a week between the administration of the first and the second tests.

Results

The correlations among the three mental ability tests were:

---

[1] Adapted and condensed from a dissertation submitted in partial fulfillment of the requirements for the Ed.D. degree at Teachers College, Columbia University, in 1955. Part of the data were collected in connection with the USPHS Project M-712 under the direction of Professor Seymour B. Sarason of Yale University.

[2] The questionnaire used was developed under the direction of Professor Sarason of Yale.

Table 1

Major Differences of Three Group Mental Ability Tests

Factor	Difference		
	Otis-Alpha	Davis-Eells	Otis-Beta
Test Items	variety of problems include geometric, pictorial, and language ability items	real-life problems, personal oriented	school achievement
Style	semiacademic	semihumorous	academic
Method of Administration	semiformal test atmosphere	game atmosphere stressed	formal test atmosphere
Instructions	little provision for praise and reassurance	substantial provision for praise and reassurance	little provision for praise and reassurance
Format	free of reading	free of reading	reading required
	small, clear pictures	large, clear pictures	print
	picture or symbol marked	large space for mark next to test item	small space on separate sheet for mark
	booklet form	booklet form	3 sheets fold like accordian
Provision for Practice	short period preceding each of two sittings	substantial periods preceding several sections	short period preceding the one sitting
Time Limit	40–50 minutes	100–120 minutes	30 minutes
Speed Factor	emphasis tends to be on power although some don't have time to try all of the items	emphasis upon power, each pupil could attempt every item	speed important, no pupil could attempt every item

Otis-Alpha vs. Davis-Eells	.59
Otis-Alpha vs. Otis-Beta	.55
Davis-Eells vs. Otis-Beta	.49

These correlations are significant at the one per cent level of probability.

The correlations among the tests and test-anxiety scores were:

Otis-Alpha vs. Text Anxiety	− .28
Otis-Beta vs. Test Anxiety	− .24
Davis-Eells vs. Test Anxiety	− .14

These correlations are significant at the five per cent level. The significance of the differences between the Test Anxiety and the intelligence test correlations was found by the use of a formula developed by Hotelling and given by Walker and Lev (9, p. 257). In this formula the correlations among the tests are taken into account. The difference between the Test Anxiety vs. Otis-Beta r of − .24 and the Test Anxiety vs. Davis-Eells r of − .14 was not found to be significant, although the results are in the predicted direction ($p = .10$). The difference between the Test Anxiety vs. Otis-Alpha r of − .28 and the Test Anxiety vs. Davis-Eells r of − .14 was found to be significant at the five per cent level of probability. It appears that the first hypothesis receives support from these results although it is most clear in terms of the Otis-Alpha vs. Davis-Eells data.

Table 2 shows that the "low" test-anxious make higher mental ability indices than the "high" test-anxious regardless of the order of test administration and regardless of the test instrument. There is also a tendency for the "High Anxious" to do better on the test taken second than on the test taken first. However, an analysis of variance of the data reported in Table 2 shows no significant variation in the size of the difference between the Davis-Eells and Otis-Beta indices because of: (*a*) the effect of the extent of test anxiety; or (*b*) the effect of the order in which tests were taken; or (*c*) the effect of the interaction of these variables. It should be pointed out, however, that the order of test administration was the greatest source of variation of the several sources investigated. The data in Table 2 clearly do not allow one to accept Hypothesis 2, although the fact that the trends are in the predicted direction suggests that consideration of other factors in future studies may give more positive results. It might be pointed out in this connection that in studies of test anxiety with college students (2, 4) and children there is evidence that the "High Anxious" subjects tend to improve rather markedly in the course of learning.

Table 2

Mean and Variance of Mental Ability Test Scores of High and Low Test-Anxious Pupils Within Order of Test Administration ($N = 33$ in each group)

	Pupil test anxiety			
	Lower 31%		Upper 31%	
Order of administration	$\bar{X}$	s^2	$\bar{X}$	s^2
Otis-Beta Administered				
First				
Otis-Beta Score	5.5	3.9	4.0	2.2
Davis-Eells Score	5.4	3.8	4.4	2.4
Davis-Eells Administered				
First				
Otis-Beta Score	5.0	3.1	3.9	4.2
Davis-Eells Score	4.7	3.1	3.4	3.5

Discussion

The results of the present study support the position that test anxiety may be related to and effect group test results. There is also some support for the Davis and Eells contention that their test is less influenced by anxiety than other conventional tests. Evaluation-connected experiences which involve failure or threat of failure may serve to engender personal reactions which interfere with test performance. The standardized group test situation may have threatening factors inadvertently attached to the experience by the test-maker or the examiner.

The anxiety-provoking role of the examiner has not been studied in this present investigation. The different examiners (teachers and psychologists) in this study may have influenced the results by provoking different anxious reactions in the children. It seems important to draw attention to such possibilities and to suggest that the examiner's role in effecting group test performance might warrant investigation.

Received April 9, 1956.

References

1. Davis, A., & Eells, K. *Davis-Eells test of general intelligence or problem-solving ability*. Yonkers, N. Y.: World Book Co., 1953.
2. Gordon, E. A. *Self esteem, need for achievement, test anxiety and performance*. Unpublished doctor's dissertation, Yale Univer., 1955.
3. Jersild, A. T., Goldman, B., & Loftus, J. J. A comparative study of worries of children in two school situations. *J. exp. Educ.*, 1941, **9**, 323–326.
4. Mandler, G., & Sarason, S. B. A study of anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47**, 166–173.
5. Otis, A. S. *Otis Quick-Scoring Mental Ability Tests. Alpha Test, Form B*. Yonkers, N. Y.: World Book Co., 1936.
6. Otis, A. S. *Otis Quick-Scoring Mental Ability Tests. Beta Test, Form Cm*. Yonkers, N. Y.: World Book Co., 1939.
7. Pintner, R., & Lev, J. Worries of school children. *J. genet. Psychol.*, 1940, **56**, 67–76.
8. Sarason, S. B., & Mandler, G. Some correlates of test anxiety. *J. abnorm. soc. Psychol.*, 1952, **47**, 810–817.
9. Walker, H. M., & Lev, J. *Statistical inference*. New York: Holt, 1953.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

The Relation of the Children's Manifest Anxiety Scale to the Concept of Anxiety as Used in the Clinic[1]

Robert D. Wirt and William E. Broen, Jr.[2]
University of Minnesota

Considerable research has been published in recent years indicating relationships between a construct measured by the Taylor Manifest Anxiety Scale (MAS) and other variables mainly within the scope of learning theory. Taylor recently has reviewed this work (2). More recently Castaneda, McCandless, and Palmero have adapted the MAS for use with children (CMAS) and have reported research indicating that its relationship to other variables approximates that of the adult form of the scale (1).

Because of the importance of the anxiety construct to clinical work, we investigated the relationship between the CMAS and the concept of anxiety as used in the clinic. Using the CMAS we tested 505 children in the 4th, 5th, and 6th grades in the Minneapolis Public Schools and 34 children from the same grades who had been referred to its Child Study Department. Predicting from the clinical concept of anxiety as we understand it, we hypothesized that the clinic children would score higher on the CMAS than the others. Analysis of variance failed to support this expectation. We also predicted there would be a positive relationship between anxiety as rated by psychologists after diagnostic study or as rated by teachers and scores on the CMAS (not verified); that intelligence would be inversely related to CMAS scores (not verified); and that there would be an inverse relationship between CMAS and reading achievement scores with intelligence partialled out (significant at the .04 level).

As a result of this research we feel that the CMAS (and the MAS insofar as it is comparable) although undoubtedly a valuable and predictive instrument is measuring something different than is denoted by the clinical concept of anxiety. We find no contradiction between this conclusion and work cited by Taylor (2) which shows a relationship between the MAS and other clinically used variables. We believe the common factor underlying the relationships to be one of "willingness to say deviant things about the self" or "test-taking attitude" which factor is not necessarily highly related to anxiety as the concept is clinically used. Studies showing significant relationships have emphasized *manifest* anxiety, as, of course, do the names of the scales.

As has been pointed out by others, concepts such as anxiety need much more careful denotation than they have yet been given. We are presently engaged in further research on the construct of anxiety with the hope of giving it greater specificity.

Brief Report.
Received July 30, 1956.

References

1. Castaneda, A., McCandless, B. R., & Palermo, D. C. The children's form of the manifest anxiety scale. *Child Develpm.*, 1956, 27, 317–326.
2. Taylor, Janet A. Drive theory and manifest anxiety. *Psychol. Bull.*, 1956, **53**, 303–320.

---

[1] An extended report of this study may be obtained without charge from Robert D. Wirt, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota, or for a fee from the American Documentation Institute. To obtain it from the latter source, order Document No. 5038 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

[2] Now at the University of Kansas.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

The "Alienation Syndrome" and Adler's Concept of "Distance"

H. L. Ansbacher
University of Vermont

In a paper published in this Journal and entitled "Alienation, social apperception, and ego structure," Anthony Davids (2) reported an Affect Questionnaire which he constructed in collaboration with Henry A. Murray. The questionnaire measures eight critical dispositions of personality: sociocentricity, optimism, and trust (three "positive" or socially desirable dispositions); and egocentricity, pessimism, distrust, anxiety, and resentment (five "negative" or socially undesirable dispositions). Statistical analysis of results obtained with this questionnaire from Harvard students led to the formulation of a syndrome termed "alienation," which is composed of the five negative dispositions.

The purpose of the present note is to show that the concept of "alienation" bears close resemblance to Adler's concept of "distance." Etymologically, the two terms have in common the idea of isolation from the group (like the sociologists' term "anomie"), the first denoting the feeling of being a stranger, the second, the attitude of standing apart from the rest.

Adler's "distance" manifests itself in a fourfold modus of, figuratively, moving backward, standing still, hesitating, or constructing obstacles (1, pp. 273–276). All kinds of anxiety are seen as forms of distance behavior. To anxiety so described, Adler subordinates a number of specific feelings and attitudes which bear close similarity to Davids' negative dispositions. In order to demonstrate this, we shall insert in the following quotation from Adler the corresponding descriptive terms from Davids.

> Once a person has acquired the attitude of running away from the difficulties of life, this attitude may be greatly strengthened and safeguarded by the addition of anxiety (Davids' *anxiety*). . . . It embitters (Davids' *resentment*) his life to a marked degree, renders him quite unsuited for making contacts, and thus also for laying the basis for a peaceful life and for fruitful accomplishments. . . . We shall meet again the well-known type who feels himself forced by necessity to think more of himself and consequently has little left for his fellow man (Davids' *egocentricity*). As though in confirmation of this view one often finds that such people are inclined to think of the past or of death, which has about the same effect. . . . Or they emphasize that all is vanity, after all, that life is short, and one cannot know what will happen (Davids' *pessimism*)" (1, pp. 276–277).

This quotation permitted the insertion of all of Davids' negative dispositions except *distrust,* which is found in Adler under the heading of Oversensitivity, one of the characteristics of the neurotically disposed individual. "Distrust in his own strength which awakens doubt about everything; distrust in others which is socially disturbing and disrupts every group—these at times characterize the oversensitive patient" (1, p. 290).

With these similarities, it is understandable that the further findings of Davids are in accordance with Adlerian theory. This is true especially for the finding "that individuals who view the world and their relationship to it in a negative (i.e., alienating) light have a weak ego structure (i.e., poor personal and social adjustment) as estimated by experienced clinical judgment" (2, p. 27). Adler had started with the observation that poor adjustment is characterized by the syndrome of distance which is the resultant of increased inferiority feelings and lack of social interest.

Davids developed the Affect Questionnaire to investigate relations between degree of alienation, social apperception, and ego structure, meaning ego strength vs. ego weakness. Not only by the formulation of the concept of alienation, but also by this purpose, is Davids' investigation in line with Adlerian theory. The relationship, according to Adler, would be: The maladjusted (alienated) individual is one who, due to increased inferiority feelings (ego weakness) and concomitant lack of social interest, apperceives the world as a hostile place. "Schema of apperception" is a term frequently employed by Adler. Specifically, the content of the schema of apperception of the maladjusted would be: They "feel and live as though they were in an enemy country. Fighting, hesitating, stopping, escaping, much more occupied with their own persons than with others, they are therefore selfish, inconsiderate, lacking in social interest, courage, and self-confidence because they fear defeat more than they desire success" (1, p. 118).

Altogether, the paper by Davids can well be taken as experimental support and extension of Adlerian theory.

Received February 23, 1956.

References

1. Ansbacher, H. L., & Ansbacher, Rowena R. (Eds.) *The individual psychology of Alfred Adler in selections from his writings.* New York: Basic Books, 1956.
2. Davids, A. Alienation, social apperception, and ego structure. *J. consult. Psychol.*, 1955, **19**, 21–27.

Journal of Consulting Psychology
Vol. 20, No. 6, 1956

Some Comments on Ansbacher's Note on the "Alienation Syndrome" and Adler's Concept of "Distance"

Anthony Davids
Brown University and Bradley Home

It is indeed gratifying when another investigator is able to make use of one's findings in building a bridge between empirical data and theory. This is especially true when the data were not gathered with a view toward supporting or refuting hypotheses derived from the particular theory with which they are eventually linked. The fact that the theory in question, that of Adlerian psychology, is not particularly new on the psychological scene, coupled with the fact that this theory has been submitted to very little experimental verification in controlled experiments, makes it seem all the more noteworthy that Ansbacher (2) was able to use our recent experimental evidence to support Adler's theoretical statements.

First, let me make a few comments on the dispositions which form the personality syndrome of alienation. On numerous occasions I have been asked why we chose to work with these particular eight dispositions of personality. That is, why did we include just these eight dispositions in our researches? Murray and I were first faced with the problem of choice of attributes to study when we embarked upon the task of constructing an auditory projective technique (5) several years ago. We wanted to include in this auditory test spoken statements and phrases indicative of certain critical dispositions of personality. By means of selective perception, apperception, accentuation, and retention of the spoken statements we hoped to secure a measure of the extent to which the perceiver possessed the various dispositions in his own personality. In our original pilot studies we experimented with many variables in addition to those which have now come to be regarded as forming an "alienation syndrome." For example, we employed other dispositions or attributes, such as self-pity, gratitude, activity, passivity, and several others.

On the basis of these pilot studies, we found that it was difficult to differentiate sharply among certain of the dispositions. That is, we found that both in writing statements to represent certain dispositions and in attempting to score subjects' responses, certain dispositions tended to overlap and some seemed not to be assessed with as much validity as others. Consequently, on the basis of practical considerations and preliminary studies, we decided to focus on the eight dispositions which are currently measured by both the auditory projective technique (5) and the Affect Questionnaire (4).

It should be pointed out, however, that deciding upon these dispositions was not merely an empirical or a theoretical development. These dispositions were proposed by Murray as being of critical importance based on his theory of personality dynamics and on his years of intensive investigation of normal and abnormal personalities. If one looks into Murray's contributions to the field of personality assessment (e.g., 7, 8, 9) it is evident that these eight dispositions, and several others, have long played a prominent role in his theorizing and his empirical investigations. Thus, the selection of these eight dispositions for study was based upon Murray's wisdom, experience, and clinical knowledge about traits or dispositions that are of critical importance in the dynamics of personality.

Turning now to consideration of the Affect Questionnaire, it was constructed originally as an instrument against which to validate findings derived from the auditory projective test. This technique attempts to assess the relative strength of the eight dispositions in subjects' personalities by having them indicate their degree of agreement or disagreement with statements designed to represent the various dispositions. It was found that two experienced clinical psychologists could agree perfectly in their classification of the statements as indicative of the different dispositions. When the Affect Questionnaire (which consists of 10 statements representative of each of the eight dispositions, making a total of 80 items) was administered to groups of subjects, it was found that there were highly significant intercorrelations among the five negative dispositions, and that scores on these dispositions were negatively correlated with scores on each of the three positive dispositions. Thus, on the basis of these empirical findings (4), which were supported by unpublished findings obtained with additional groups of both normal and maladjusted subjects, we labeled the interrelated negative dispositions as forming an "alienation syndrome."

From the brief description of the selection of dispositions for study and the empirical findings which led to formulation of the syndrome, it is quite evident that the link with, or similarity to, Adler's concept of "distance" was not readily apparent to us, nor were our researches motivated by Adlerian theory. Thus, it seems quite striking that our empirical findings fit so neatly with Adler's theoretical formulations, and that these dispositions do, in fact, tend to coexist in the personality just as Adler believed on the basis of his theoretical convictions and his clinical experience.

Today much research attention is being devoted to submitting to experimental test many of the theoretical notions put forth by Freud on the basis of his clinical experience. However, this does not seem to be the case with Adlerian theory. It is, therefore, especially interesting to find that an experiment such as ours, which fits into the currently popular area of research termed "social perception" and which uses experimental techniques and statistical procedures, fits so nicely with hypotheses that would be generated by Adlerian theory. In other words, if we had taken as our point of departure the theoretical notions and statements advanced by Adler, and had designed our researches specifically to test propositions derived from his theory, we would have confirmed them in a well-controlled laboratory investigation employing modern statistical procedures. Since there is, undoubtedly, much of value in Adler's theory (3), and since many features or derivatives of Adlerian theory are put to use in present-day clinical practice (1, 6, 10), it seems of real value that Ansbacher was able to perceive the link between our seemingly unrelated experiment and Adler's theoretical concepts.

Received February 23, 1956.

References

1. Ansbacher, H. L. "Neo-Freudian" or "Neo-Adlerian"? Report on a survey conducted among members of the American Psychoanalytic Association. *Amer. Psychologist,* 1953, **8,** 165–166.
2. Ansbacher, H. L. The "alienation syndrome" and Adler's concept of "distance." *J. consult. Psychol.,* 1956, **20,** 483–484.
3. Ansbacher, H. L., & Ansbacher, Rowena R. (Eds.) *The individual psychology of Alfred Adler in selections from his writings.* New York: Basic Books, 1956.
4. Davids, A. Alienation, social apperception, and ego structure. *J. consult. Psychol.,* 1955, **19,** 21–27.
5. Davids, A., & Murray, H. A. Preliminary appraisal of an auditory projective technique for studying personality and cognition. *Amer. J. Orthopsychiat.,* 1955, **25,** 543–554.
6. Munroe, Ruth L. *Schools of psychoanalytic thought.* New York: Dryden, 1955.
7. Murray, H. A. *Explorations in personality.* New York: Oxford Univer. Press, 1938.
8. Murray, H. A. *Thematic Apperception Test Manual.* Cambridge: Harvard Univer. Press, 1943.
9. Murray, H. A., & Morgan, Christiana D. A clinical study of sentiments. *Genet. Psychol. Monogr.,* 1945, 32, 3–149.
10. Thompson, Clara. *Psychoanalysis: evolution and development.* New York: Hermitage House, 1950.

PSYCHOLOGICAL TEST REVIEWS

Blum, Gerald S. *The Blacky Pictures.* New York: Psychological Corp., 1950.

Reviewed by

Samuel J. Beck
University of Chicago

A homely, obstreperous, and by now not unloved canine, who plays the title—and perhaps a hero's—role, this is Blacky of the Blacky test. Sired by psychoanalysis out of TAT he was created *ad hoc:* to test out psychoanalytic theory concerning psychosexual development, *Videlicet,* the title of Blum's monograph (2). It is a bold ingress into that most fluid area of psychology—depth psychology—which Blum undertakes, one more specific in purpose and design than those which projective tests have so far essayed. These, pioneers that they are, opening up a totally new field, have been engaged in attempting some broad estimate of the new territory, appraising its potential yield in terms of fresh knowledge concerning the human personality or, at the least, fresh ways of studying it. They have been treading warily, with some trembling at the shapes and sights unholy lurking in the scientific jungle that surrounds the regions of the human personality. However, as at other frontiers, the pioneers are followed by the more intensive cultivators. These latter can stake out, each his special acre, and so a method like that of the projective tests advances, by new experiment.

Blum has devised a test *ad hoc* for psychoanalytic theory by the ingenious technique which activates certain of the psychoanalytically formulated unconscious traits, at the same time that it provides the defenses enabling the patient to liberate and communicate these forbidden trends—the defenses of identification and displacing. I am not using any of the space allotted to me for this review in describing the test. He who has not yet made the acquaintance of the very human Blacky is referred to his creator's monograph (2). In carrying on his experiment, Blum used, as his point of departure, thirteen "dimensions" in psychoanalytic theory. Anyone even superficially read in this field will recognize the terms. His results, in University of California students, showed statistically significant differences between the sexes, statistically significant relationships between dimensions, and evidence in the writings of Freud and of Fernichel (he quotes the text from both) showing an impressive consistency between Blum's significant findings and psychoanalytic theory. He arrives at the cautious conclusion that "some aspects of psychoanalytic theory have demonstrable validity."

In the few years since he published, his test and findings have been subjected to fairly extensive research, not surprising what with the dominance of both psychoanalysis and projective instruments on the current psychological scene. Within a year, Seward criticized the exclusion of statistically insignificant but relevant data in Blum's report, and also pointed to some likely logical flaws (11). Michal-Smith, Hammer, and Spitz found their Blacky test analysis consistent with indication of Oedipal intensity in a nine-year-old boy (8). Paranoids were studied by Aronson (1) and he reported partial confirmation of Blum's results. Other investigations that have come to my attention (I am making no pretense here to an exhaustive survey of the literature) are: Blum and Kaufman (4) in 14 male patients with peptic ulcer; McNeil and Blum (7) relating handwriting signs with the psychosexual dimensions; and Ellis (6), who, in a patient with some two hundred hours of psychoanalysis, had the test rated by 22 other psychologists. In a survey published in 1952 by Blum and Hunt (3), they reported also an investigation in stutterers (Merchant); sexual offenders (Lindner); behavior in small groups (Swanson). A more recent investigation is that of Molish, Lyon, and Briggs (9), who compared Blacky test findings in 1947 naval recruits found "not maladjusted" by the Navy's neuropsychiatry service, with 390 recruits who were rejected as unsuitable. These investigators, while alert to the psychoanalytic implications of their results, were especially interested in the test's potency for differentiating their two groups. An inference from their research is that, whatever the agreements with psychoanalytic theory, the Blacky has its value in practical application, as derived from empirical data.

The test has thus a differentiating potency, and without a doubt it has the projective instrument value of opening a window to latent character traits. At the same time results are suspect, as I see it, by reason of a major fallacy in the technique of treating the data. This consists in the method of "scoring" each story and the other criteria as "strong" and the various gradations of "strong." The judgments are made by each examiner, without recourse to any frame of reference. Scorings must, therefore, be a function of each scorer. They can, I will concede, be made reliable as between two investigators who have worked together closely and have learned the same signs and language, in recognizing and naming the psychological traits or "dimensions." Except between two such colleagues, or for some one examiner, results are only spuriously comparable. The degree of error to which the procedure is subject is seen in Blum's own monograph: a denial of a response is construed as a repression, and since a repression involves strong emotion, the response is scored "strong" (**2**, p. 34). To construe denial as repression—which is a hypothesis and not a datum—is interpretation. Thus a quantitative profile is developed and is reported as *S*'s production although it is actually the examiner's interpretation of *S*'s production.

This criticism touches on what is the Achilles' heel of each projective test. It dictates the need, therefore, of exploiting the objectivity potential in the test. The needed technique is that of patterning out its data into impersonal structure to be described operationally, and so creating some stable frames of reference by which the productions of any *S* can be judged, if not measured. Some of the projective tests do lend themselves more easily to such impersonal structuralization than others. But to the extent that the instrument falls short of such objectivity, it is not a test but a two-person situation, two persons—*S* and examiner, each of whom may influence the results. Schafer (**10**) has been especially emphatic concerning the interpersonal relation between examiner and subject, and the possible role of the examiner's personality in fashioning the final test picture. Brunswik writing from another theoretical viewpoint, arrives at a similar conclusion. "Since in any testing procedure the examiner constitutes part of the external stimulus situation, representative design demands that examiners should also be sampled," and he refers to the "double jeopardy of generalization" (**5**, p. 131).

An important implication follows from these positions of Schafer and of Brunswik: projective test investigations will need to find some method of keeping the examiner factor constant, if reports by two or more persons are to be comparable. The problem becomes at this point the baby of the theorists and logicians of science. This circumstance does not, however, ease the burden of the psychologists searching personality by means of projective tests. The fact is they have so far been arriving at righ answers, even if they do not know how. Their answers are bound to be that much more accurate as they are based in sound method, and as they take pains to devise such. This is where I see Blacky and his growing pains at the present. The test has the promise warranting its developing. In spite of the fallacy in the scoring procedure, its agreement with clinical theory is logically sound. It can serve, further, as an example of other *ad hoc* techniques which can be devised to explore psychoanalytic theory.

References

1. Aronson, M. L. A study of the Freudian theory of paranoia by means of the Blacky pictures. *J. proj. Tech.*, 1953, **17**, 3–19.
2. Blum, G. S. A study of the psychoanalytic theory of psychosexual development. *Genet. Psychol. Monogr.*, 1949, **39**, 3–99.
3. Blum, G. S., & Hunt, H. F. The validity of the Blacky pictures. *Psychol. Bull.*, 1952, **49**, 238–250.
4. Blum, G. S., & Kaufman, Jewel B. Two patterns of personality dynamics in male peptic ulcer patients as suggested by responses to the Blacky pictures. *J. clin. Psychol.*, 1952, **8**, 273–278.
5. Brunswik, E. *Perception and representative design of psychological experiments.* Berkeley, Calif.: Univer. of California Press, 1956.
6. Ellis, A. The Blacky test used with a psychoanalytic patient. *J. clin. Psychol.*, 1953, **9**, 167–172.
7. McNeil, E., & Blum, G. S. Handwriting and psychosexual dimensions of personality. *J. proj. Tech.*, 1952, **16**, 476–484.
8. Michal-Smith, H., Hammer, E., & Spitz, H. The use of the Blacky pictures with a child whose Oedipal desires are close to consciousness. *J. clin. Psychol.*, 1951, **7**, 280–282.
9. Molish, H. B., Lyon, B., & Briggs, D. L. Character structure of adjusted and maladjusted naval recruits as measured by the Blacky pictures. *Amer. J. Orthopsychiat.*, 1954, **24**, 164–174.
10. Schafer, R. *Psychoanalytic interpretation in Rorschach testing.* New York: Grune & Stratton, 1954.
11. Seward, J. P. Psychoanalysis, deductive method, and the Blacky test. *J. abnorm. soc. Psychol.*, 1950, **45**, 529–535.